Title
Ordering, Measurement, and Ordinal Measurement: A Pragmatic Perspective

Author
Torres Irribarra, David

Publication Date
2016

Peer reviewed|Thesis/dissertation

Ordering, Measurement, and Ordinal Measurement: A Pragmatic Perspective

by

David Torres Irribarra

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Nicholas Jewell

Summer 2016

Ordering, Measurement, and Ordinal Measurement: A Pragmatic Perspective

Abstract

Ordering, Measurement, and Ordinal Measurement: A Pragmatic Perspective

by

David Torres Irribarra

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This dissertation encompasses three papers that touch on the topics of the definition of measurement, the possibility of ordinal measurement and the application of an ordinal psychometric model.

The first paper, *A Pragmatic Perspective on Measurement*, addresses the issue of the definition of measurement, presenting a conceptualization of the practice of measurement from the perspective of the Pragmatic tradition in philosophy (Bacon, 2012; James, 1907/1995; Rorty, 1999). In the spirit that Pragmatic approach, this definition is put forward as a tool aimed at gauging measurement claims in terms of their usefulness, such that it can contribute to a better understanding between researchers, practitioners and users of measurement. The paper discusses central ideas of the Pragmatic tradition and reviews the main measurement traditions in the social sciences before making a case for a Pragmatic Perspective of Measurement and offering a definition of measurement based on that approach. The proposed definition attempts to achieve this by bringing to the foreground a conception of measurement according to which: (a) order and classification are part of measurement as well as quantification, (b) the model of the attribute that underlies a measurement is central to designing and interpreting measurements, (c) attributes need not be considered as natural kinds or universals, and (d) the purpose for which the measurement was developed informs us about the scope of its utility, both to judge its success as well as limiting the inferences that can be supported by it.

The second paper, *Categorization, Ordering and Quantification: Selecting a Latent Variable Model by Comparing Latent Structures*, is a joint work with Ronli Diakow that proposes a model selection framework for identifying the kind of latent structure—classificatory, ordinal, or quantitative—that best describes a dataset. The framework and its rationale for successive comparison of models outlined in this paper offers a blue-print to directly addressing issues that so far are largely thought to only be examinable under the representational theory of measurement, namely, the empirical identification of ordinal and quantitative structure in an attribute. The possibility of analyzing them under a latent variable framework would allow the critical examination of the assumptions regarding the latent structure that can be supported based on the data and, more generally, to question and revise our assumptions regarding the structure that we ascribe to the relevant attributes that we study.

The last paper builds on the previous papers—which present the arguments that measurement can be ordinal, and that it is possible to identify cases when a model that assumes an ordinal latent structure is better suited to a dataset— by introducing *The Ordered Mixture Linear Logistic Test Model* (OM-LLTM), an explanatory item response theory model conceived for ordinal measurement. The OM-LLTM is a model suited to take advantage of the cases when we have a theory that describes a relevant attribute in terms of a set of ordered performance levels, and we construct our assessment instruments according to that theory. The OM-LLTM assumes respondents are grouped in ordered latent classes where the probability of correctly answering an assessment task is a function not only of the class membership of the respondent, but also of item features that—according to the theory—determine the difficulty of the task. This model is a combination of the Linear Logistic Test Model (LLTM; Fischer, 1973) and Ordered Latent Class Analysis (OLCA; Croon, 1990). The OM-LLTM can also be considered an ordered extension of the mixture LLTM developed by Mislevy and Verhelst (1990).

The combination of these two models will allow researchers and practitioners to model student proficiency according to *explanatory* (De Boeck & Wilson, 2004) models expressed through the LLTM part of the model, while providing simple and interpretable results in terms of ranked performance groups, through the OLCA part of the model. Accordingly, the OM-LLTM offers both a simple, *coarser*, interpretation of the respondent classes according to overall proficiency and also an explanatory

interpretation in terms of the specific item features; where the former interpretation lends itself for use in context where summative assessments are needed and the latter is more appropriate when diagnostic information is required.

To those who believed in me, even when I did not believe myself;
and to those who supported me, even when I could not reciprocate.

Contents

List of Figures

List of Tables

Acknowledgments

This has been a long, long journey; one that I quite simply could not have attempted, much less finished, on my own. Faced with the opportunity of thanking all the people that have made it possible for me to complete my dissertation, I realize that words simply cannot convey the extent of my gratitude, nor is the space enough to list them all. Despite this, I would like to express my thanks—more or less chronologically—to at least some of them:

To my family and especially Virginia, my mother, whose guidance shaped me, whose faith in me gave me courage, and whose support and love have defined me.

To Carlos and Roberto, my mentors in my undergraduate years, I would not have dreamed of doing this if not for your encouragement, and I could not have done it without the tools each of you gave me.

To Anita, my wife, who with kindness taught much of the mathematics I needed to understand psychometrics, and whose endless patience, care, and love I will never be able to repay.

To Ronli, friend, officemate, and co-author, for engaging with me in discussion on any and all topics while challenging on me in every one of them; half my doctoral studies consisted in just learning from you.

To Rebecca, dear friend and avowed board game nemesis, for accompanying me during some of what seemed the longest years in this dissertation. I will be forever grateful for your kindness, and for you finding that damned bug in my code.

To Andy, friend and colleague, for all the drafts that you read and commented on, and for your warmth and thoughtfulness both inside and outside the academic world.

To Mark, my mentor, first and foremost, thank you for your incredible patience in each and every one of our disagreements, as I know I was—and still am—a difficult student. Your example and guidance will accompany me in all my future endeavors.

To Sophia, Nick, Erin, Patrick, Bruce, Geoff, and Maryl, my professors, I am truly lucky to have had the opportunity of learning from you.

To all the QMEs and associated post-docs, past and present, especially Amy D., Coulter, Beth, Perman, Tina, Katherine, and Amy A. Our conversations in the QME lab, our discussion during study group, and all the reveling during happy hour made the program feel like home.

Chapter 1

Introduction

Measurement has been for a long time a ubiquitous activity, one that is associated with the production of valuable information that we can use in order to make decisions in many domains of our lives (Duncan, 1984). The prototypical associations to what it means to measure an attribute have historically been connected with physical attributes, such that we tend to think of rulers that measure distance and scales that measure weight as prime examples of this practice. However, in the last century or so the social sciences have striven to actively engage in this activity, attempting to *measure* a wide variety of social and psychological concepts (Michell, 1999; Woolf, 1961). This endeavour has been successful at the very least to the extent that it is no longer strange to hear, and accept, that psychologists can measure a person's intelligence or that educators can measure students' level of mathematical knowledge or writing ability.

However, despite the successful adoption of the idea that measurement is not activity limited to mass, distance and other physical attributes, but also an activity that can be conducted on social and psychological attributes, it is not always clear how this jump from the physical to the psycho-social has changed the way in which we understand measurement. Do we mean exactly the same thing when we claim to measure length as when we claim to measure anxiety? If not, what is the difference? Should we trust in thermometers in the same way in which we trust personality tests? Are there limits to what can be measured? These are central questions that raise issues regarding the possibility of conducting measurement at all in the social sciences, the way in which the measurement procedures and outcomes can be interpreted, and the

role that measurement plays in the development of social and psychological theories. Many scholars have raised these and other issues, often criticizing the use of the concept of measurement in the social sciences, and the reticence of social scientists to cope with these theoretical and philosophical questions about the definition and practice of measurement (Boring, 1920; Borsboom, 2005; Cliff, 1992; H. M. Johnson, 1936; Michell, 2008b; Schönemann, 1994; Trendler, 2009).

These questions about the very possibility of measurement in the social sciences exist at the same time that the practice is widely accepted in society, serving as the basis for personnel selection, educational placement, and international comparisons. In this context, it is not an option to simply expect researchers, technicians, and consumers to stop what they are doing until we achieve a coherent and widely accepted basis for conducting psycho-social measurements. At the same time, it is not tenable to continue trusting implicitly in any claim of measurement of psycho-social concepts to the extent that some of these basic questions regarding the meaning and validity of the practice remain unanswered. This situation brings to mind Neurath's famous allegory (1973):

> "We are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction." (p. 199)

The three papers in this dissertation are presented in the spirit of a gradual effort for reconsidering—or reconstructing as the allegory has it—the ship of measurement in the social sciences.

1.1 Defining Measurement

The first paper, *A Pragmatic Perspective on Measurement*, addresses the question for the definition of measurement, presenting a conceptualization of the practice of measurement from the perspective of the Pragmatic tradition in philosophy (Bacon, 2012; James, 1907/1995; Rorty, 1999).

The paper begins with the formulation of key questions regarding the scope and limits of measurement, emphasizing that if the trust that the public places on measures in the social sciences relies on their connection to the notion of measurement in the physical sciences, then the clarification of the similarities and differences between measurement in the physical and the social realms is of central importance to adequately contextualize their relative advantages and limitations.

This paper reviews some of the most influential theories of measurement such as the "classical view" of measurement (Michell, 2005), operationalism (Bridgman, 1927), the representational theory of measurement (Krantz, Luce, Suppes, & Tversky, 1971/2007; Pfanzagl, 1968), as well as more methodological perspectives arising from the practice of researchers in the social sciences, such as the latent variable perspective (Lazarsfeld & Henry, 1968; Lord & Novick, 1968; Rasch, 1960/1980), and in the physical sciences and engineering, represented by metrology. This overview illustrates that the concept of measurement, and that of quantitative methods, is currently being used across the board in ways that do not necessarily conform to traditional (i.e. classical) definitions of measurement, pushing the boundaries of what constitutes our technical understanding of it. Moreover, what constitutes a technical understanding of measurement, and the theoretical commitments that it entails, will vary in different areas. In this context, disagreement on what is constitutive of measurement is bound to appear.

Pragmatism offers here the advantage of being flexible and fallibilist, encouraging us to abandon the pursuit of a timeless and perfect definition that attempts to establish decontextualized/definitive demarcation criteria for what is truly measurement.

From this Pragmatic Perspective of Measurement, I propose to understand measurement as:

> (a) an activity of classification, ordination, or quantification of a set of elements (b) according to a model (c) of a relevant attribute (d) in service of a larger goal.

In the spirit of a Pragmatic approach, this definition is put forward as a tool aimed at gauging measurement claims in terms of their usefulness, such that it can contribute to a better understanding between researchers, practitioners and users of measurement. This definition attempts to achieve this by bringing to the foreground

a conception of measurement that encompasses order and classification as well as quantification (point *a* in the definition), that the model of the attribute that underlies a measurement is central to designing and interpreting measurements (point *b*), that attributes need not be considered as natural kinds or universals (point *c*), and that the purpose for which the measurement was developed informs us about the scope of its utility, both to judge its success as well as limiting the inferences that can be supported by it (point *d*).

1.2 Identifying Classes, Orders, and Quantities

The second paper, *Categorization, Ordering and Quantification: Selecting a Latent Variable Model by Comparing Latent Structures*, is a joint work with Ronli Diakow that proposes a model selection framework for identifying the kind of latent structure—classificatory, ordinal, or quantitative—that best describes a dataset.

As part of the debate regarding the definition and possibility of conducting measurement in the social sciences, Michell (1999, 2008b) has emphatically criticized the lack of examination of the assumption that psycho-social variables are quantitative. The skepticism regarding the possibility of modeling quantitatively psychological properties, known as the *quantity objection*, can be traced back at least to the beginning of the 20th century (Boring, 1921), and has been echoed through the years by multiple scholars (H. M. Johnson, 1936; Michell, 2008b; Schönemann, 1994; Trendler, 2009). The criticism of these authors has pointed towards the failure of social scientist to adopt the methods developed in the representational theory of measurement for testing the presence of a quantitative structure in the data, arguing that so far this is the only well established method for conducting such test in the social sciences (Michell, 2008b). More generally, the unexamined acceptance of the structure of an attribute of interest—classificatory, ordinal, or quantitative—without evidence to support such a decision is problematic, as it implies the acceptance of a specific way of construing and making inferences about the relevant attributes that are being studied, which in the social sciences are often persons.

This paper introduces a model selection framework that would directly address this concern over the need for and evidence-based approach to the selection of a latent structure, taking advantage of a variety of latent variable models to collect

evidence supporting the selection of that structure. In order to do so, we rely on the differences between quantitative, ordered, and qualitative latent structures, which can be represented by a series of the latent variable models discussed in the previous section that progressively assume a more constrained latent structure.

In order to capture the range of possible structures for the latent variable we will rely on six models:

(i) The Unconstrained Latent Class Model (UN).
(ii) The Ordered Latent Class Model with Class Monotonicty (MON).
(iii) The Ordered Latent Class Model with Invariant Item Ordering (IIO).
(iv) The Ordered Latent Class Model with Double Monotonicity (DM).
(v) The Located Latent Class Model or Latent Class Rasch Model (LCR).
(vi) The Rasch Model (RM).

The identification of this structure is achieved by comparing the fit between the six aforementioned models and selecting the model that offers the best relative fit to the data. Using the model identified with this framework would allow us to make tenable claims about the kinds of inter-individual differences that can be detected in the data.

The overall rationale for the comparison outlined in this paper offers a blue-print to directly addressing issues that so far are thought to only be examinable under the representational theory of measurement. The possibility of examining them under a latent variable framework would allow the critical examination of the assumptions regarding the latent structure that can be supported based on the data and, more generally, to question and revise our assumptions regarding the structure that we ascribe to the relevant attributes that we study.

1.3 Embedding Theory Into Ordinal Measurement

The previous papers have made the case that (a) measurement can be ordinal, and (b) that it is possible to identify cases when a model that assumes an ordinal latent structure is better suited to a dataset. This last paper, *The Ordered Mixture Linear Logistic Test Model* (OM-LLTM), introduces this item response theory model suitable for ordinal measurement.

The OM-LLTM is a model suited to take advantage of the cases when we have a theory that describes a relevant attribute in terms of a set of ordered performance levels, and we construct our assessment instruments according to that theory. The OM-LLTM assumes respondents are grouped in ordered latent classes where the probability of correctly answering an assessment task is a function not only of the class membership of the respondent, but also of item features that—according to the theory—determine the difficulty of the task.

This model is a combination of the Linear Logistic Test Model (LLTM; Fischer, 1973) and the Ordered Latent Class Model (OLCA; Croon, 1990). The OM-LLTM can also be considered an ordered extension of the mixture LLTM developed by Mislevy and Verhelst (1990).

The combination of these two models will allow researchers and practitioners to model student proficiency according to *explanatory* (De Boeck & Wilson, 2004) models expressed through the LLTM part of the model, while providing simple and interpretable results in terms of ranked performance groups, through the OLCA part of the model.

Ordered performance groups are commonly adopted in educational contexts by policy makers and assessment developers to characterize different performances in learning situations. "Below basic," "basic," "proficient," "advanced," are just one example of the type of ordered levels commonly used in school settings.

Examples of this widespread practice can be found in the performance levels used to communicate results in international tests such as PISA (OECD, 2007), TIMSS (National Center for Education Statistics, 2009), the NAEP achievement levels (Bourque, 2009) and the myriad of performance classifications used to determine the percentage of "proficient" students (Perie, 2008).

As is the case with many of the aforementioned examples, the construction of performance levels often arises as a product of practical necessities, such as communicating assessment results to stakeholders with non-technical backgrounds or to comply with legal requirements such as the ones established by NCLB. However, these performance levels can also be motivated as part of attempts to develop cognitive and instructional theories (Mislevy, 1996; National Research Council, 2001; Wilson, 2005), representing hypotheses about the way in which students' understanding of a domain develops, such as the recent movement that promotes the characterization of

developmental pathways in terms of *Learning Progressions* (Lehrer, Kim, Ayers, & Wilson, 2014; National Research Council, 2006; C. Smith, Wiser, Anderson, & Krajcik, 2006; Wilson, 2009).

In the cases where such a theory is available, and (a) it specifies the features that should govern the difficulty that students have when answering questions in a given domain, and (b) that theory has been used to inform the construction of assessment instruments, the OM-LLTM offers both a simple, *coarser*, interpretation of the respondent classes according to overall proficiency and also an explanatory interpretation in terms of the specific item features; where the former interpretation lends itself for use in context where summative assessments are needed and the latter is more appropriate when diagnostic information is required.

Chapter 2

A Pragmatic Perspective on Measurement

2.1 Introduction

Let me start with an imaginary dialog between a researcher who recently finished a study on anxiety and a skeptic.

> **Researcher** As a part of my study I measured anxiety.
>
> **Skeptic** How exactly did you manage to do that?
>
> **Researcher** Well, I designed a questionnaire with five Likert-type questions, collected a lot of data, analyzed the results to ensure the quality of the instrument, and, after checking that everything was in order, I calculated an estimate of anxiety for each participant in the study.

The researcher seems to be following usual practice in psychology and the social sciences[1], but the question remains, did the researcher *really* measure anxiety? When he says "I measured anxiety," the skeptic wonders whether she can place similar trust as she would place in the results of a speedometer or a thermometer gauge, or similar instances of measurement that have gained acceptance and respect broadly in our society. After all, as Van Fraassen (2008) points out:

[1] See for example Knapp and Mueller (2010) for an example on recommended guidelines for reviewers of manuscripts presenting measurement instruments.

> The term "measurement" is an endorsing term. If we call something a measurement, we imply that there is something correct or valuable in the way it yields a representation (even if on this particular occasion it went wrong). Measuring is an operation by which we can produce or gather information; and here "information" too has an endorsing sense. (p. 151)

The high regard for the label of *measurement* has increased over time due to the unparalleled success of measuring and modeling of the physical world, which Wigner (1960) described as *"the unreasonable effectiveness of mathematics in the physical sciences"*[2] (p. 222), and has made it so that *"in our own time, measurement means nothing if not precision and objectivity"* (Porter, 1996, p. 23).

The social sciences have long aimed at replicating this success, and in this spirit have pursued the measurement of their concepts of interest. "Quality of life," "intelligence," "aptitude," "knowledge," "customer satisfaction," "job performance," "similarity," "utility," "prejudice," "well-being," "socio-economic status," and "social capital" are all examples of concepts that social scientists claim to have measured.

In the face of all these measurement claims, the emergence of a certain level of skepticism (cf. H. M. Johnson, 1936; Trendler, 2009) seems natural and healthy. Is there a limit at all to what can be measured? What does it mean to measure something? It is not clear how far we are willing to extend that concept.

Is a measure of the diagonal of a screen equivalent in trustworthiness to the measure of someone's mathematical ability? If the trust that the public places on measures in the social sciences relies on their connection to the notion of measurement in the physical sciences, and their results are used to make inferences and decisions about persons and public policies, then the clarification of the similarities and differences between measurement in the physical and the social realms is of central importance to adequately contextualize their relative advantages and limitations.

[2] Wigner recognizes that this effectiveness is a product of an historical development process; "It is true, of course, that physics chooses certain mathematical concepts for the formulation of the laws of nature, and surely only a fraction of all mathematical concepts is used in physics. It is true also that the concepts which were chosen were not selected arbitrarily from a listing of mathematical terms but were developed, in many if not most cases, independently by the physicist and recognized then as having been conceived before by the mathematician." (p. 229), yet he is still amazed by its results.

To address these concerns about the scope and limits of measurement, scholars throughout history have attempted to define a clear demarcation between measurement and non-measurement (Borsboom, 2005; Campbell, 1920; Mari, 2003; Michell, 1999; Savage, 1970; Stevens, 1946). Such a criterion would supposedly allow us to separate quackery and posturing from *real measurement*, by providing a way of testing which claims should enjoy the trust of the scientific community and public opinion. Some have looked for the key to defining the concept of measurement in the nature of the attributes that we purport to measure (Michell, 2005), others have emphasized the methods that we must follow to achieve it (Krantz et al., 1971/2007; Mari, 2000), and still others have even focused on its products (Dingle, 1950; Speitel, 1992). Despite their differences, most if not all these efforts attempt to answer the question: What is the *true* and conclusive definition of measurement? I, too, have found myself tempted by this longing for that definition of measurement that will finally provide us with a clear-cut criterion. However, despite the many answers that can be found in the literature, I have come to think that the problem lies in formulating the question in this manner.

In this paper, I would like to elaborate how Pragmatism, a philosophical perspective that turns away from ahistorical accounts of knowledge, embracing instead "the role of *know-how*, *social practices*, and *human agency*" (Bernstein, 2010, p. 9), can help us address questions about the definition and limits of measurement; and how in doing so it can challenge some deep-seated assumptions shared by most, if not all, of the alternative accounts of measurement cited above, so that we can think about it in a different, and hopefully more productive, way.

2.2 What is Pragmatism?

So, what do I mean when I talk about Pragmatism?

A first clarification is to distinguish what I would consider colloquial uses of the term "pragmatic" from the Pragmatist tradition that I refer to across this paper.

In its colloquial sense, "pragmatic" is oftentimes associated with *practical* as being somehow devoid of, opposed to, or at least not requiring, theory. A clear example of this usage in the measurement literature can be found in Hand (2004, p. 53)[3]:

[3]In his treatment of what he calls "pragmatic measurement" Hand links it to operationalism,

> ...the pragmatic part is purely operational, and is simply an arbitrary aspect of the process of taking the measurement—arbitrary in the sense that other measurement procedures may adopt alternative pragmatic constraints.

This is not the sense in which I use the word in this paper. I discuss how theory-building is one more of the many activities that humans engage in coping with the world, and how from a Pragmatist's perspective, theories are powerful tools in their own right. In the sense used in this paper, Pragmatism is not evoked to say "we don't really need to think about theory."

A second informal use of the word "pragmatic" is to mean *expedient* or *convenient*, and somehow limited to a short-term perspective. Short-term consequences and applications can certainly be considered, but I want to emphasize that the Pragmatist perspective described in this paper does not restrict the focus to them. In any given issue there are interesting discussions to be had regarding uses and consequences in the long- and medium- versus the short-term, but a Pragmatic attitude does not necessarily favor any of them in principle.

Finally, it is important to recognize that many people consider themselves to be practical or pragmatic. I do not pretend to be the arbiter of what is the correct use of the word or who gets to be in the "pragmatic club." However, in this paper, I refer to Pragmatism as the approach developed in the Pragmatic tradition associated in its origins with C.S. Peirce (1878/2014), William James (1907/1995), John Dewey (1929/1960, 1920/2004), and more recently with the neo-Pragmatism of Richard Rorty (1982, 1999).

As with any other tradition, Pragmatism is hardly monolithic, prompting Westbrook (2008, p.185) to characterize it "less as a well defined, tightly knit school of thought than as a loose, contentious family of thinkers who have always squabbled, and have sometimes been moved to disown one another." Throughout this paper, I try to present the common or more distinctive elements of this tradition, and when appropriate I discuss which family members' interpretations I am favoring.

That was a long preamble... so, what is Pragmatism?

---

and although he explicitly decries the confusion between them, his treatment conflates them: "the results of any [unrestricted] transformation would define an equally legitimate, pragmatic measurement procedure. Of course, it would be a different procedure... so perhaps we should give the resulting variable a different name." (p. 57-58)

2.2.1 A Brief Overview of Some Pragmatist Ideas

The Primacy of Practice

Briefly speaking, Pragmatism is a philosophical tradition that emphasizes that the meaning of an idea or concept can be characterized by the effects that it produces. This core idea was expressed by C.S. Peirce (1878/2014) in what is now known as the *Pragmatic maxim*:

> Consider what effects, which might conceivably have practical bearings, we conceive the object of our concept to have. Then, our conception of these effects is the whole of our conception of the object. (p. 31)

The maxim directs our attention toward what James (1907/1995) called the *practical cash-value* of our ideas and theories in order to help us understand them and judge their worth. On the centrality of this currency Peirce (1878/2014) indicates that "we come down to what is tangible and practical, as the root of every real distinction of thought, no matter how subtle it may be; and there is no distinction of meaning so fine as to consist in anything but a possible difference of practice" (p. 30). This emphasis is a central thread that connects Pragmatic thinkers, even though different members of the Pragmatic family may advance different interpretations of what "tangible and practical" means (Bacon, 2012).[4]

It is worth mentioning at this point that a kindred approach to meaning can also be found in Ludwig Wittgenstein's later work, where he discusses meaning as tied to use:

> The meaning of a phrase for us is characterized by the use we make of it. The meaning is not a mental accompaniment to the expression....— I want to play chess, and a man gives the white king a paper crown, leaving the use of the piece unaltered, but telling me that the crown has a meaning to him in the game, which he can't express by rules. I say:

[4]Perhaps the clearest example of a divergence on this issue occurs between Peirce and James, where the latter interpreted the notion of "practical" to include moral, aesthetic and religious concerns, while the former restricted it to experimental/observational terms (Bacon, 2012; Rescher, 2007).

> "as long as it doesn't alter the use of the piece, it hasn't what I call a meaning." (Wittgenstein, 1965, p. 65)

Even though he is not usually considered a Pragmatist, Wittgenstein's work has been actively discussed and used by leading Pragmatists such as Rorty (1979) and Putnam (1995), and the affinity of his work with some key Pragmatic ideas has been often noted and discussed (e.g., R. B. Goodman, 1998; R. Haack, 1982; Hensley, 2009; J. Margolis, 2009). Because of this affinity, I avail myself of Wittgenstein's work throughout this paper when arguing for a Pragmatic perspective of measurement.

Fallibilism and Anti-skepticism

A second central aspect of the Pragmatic tradition, going back to Peirce's article *Some Consequences of Four Incapacities* (1868), is the commitment to both fallibilism and anti-skepticism (Bacon, 2012). To be a fallibilist means accepting that any belief can be subject to revision, no matter how entrenched or central to us; to put it bluntly, it means accepting that we can always be wrong about anything. Although this might seem to be courting utter skepticism, Pragmatists explicitly distance themselves from this position, assuming an anti-skeptic position that means that "while any particular belief might, in the course of inquiry, be upset and overturned, they are unproblematically relied upon until reason to question them is given" (Bacon, 2012, p. 20). This balancing act between fallibilism and anti-skepticism is rooted in Peirce's doctrine that "reasoning should not form a chain which is no stronger than its weakest link, but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected" (Peirce, 1868, p. 141).

Bernstein (1998) argues that the impact of Peirce's doctrine expressed in his 1868 article can be seen almost a century later in the work of Sellars (1956). In his 1956 article "Empiricism and the Philosophy of Mind" Sellars states that "empirical knowledge, like its sophisticated extension, science, is rational, not because it has a *foundation* but because it is a self-correcting enterprise which can put *any* claim in jeopardy, though not *all* at once" (p. 300). The importance of the fallibilist and anti-skeptic stance for the pragmatic tradition is best summarized in the words of Putnam (1995):

> Pragmatists hold that *doubt* requires justification just as much as belief.... [and] that there are no metaphysical guarantees to be had that even

> our most firmly-held beliefs will never need revision. That one can be both fallibilistic *and* antisceptical is perhaps *the* basic insight of American Pragmatism. (p. 20-21)

Historical and Social Nature of Knowledge

Another strand of Pragmatist thought is its rejection of an "ahistorical, transcendental, or metaphysical theory of truth" (Misak, 2013, p. 3), and embracing instead "a nonfoundational self-corrective conception of human inquiry based upon an understanding of how human agents are formed by, and actively participate in shaping, normative social practices" (Bernstein, 2010, p. x).

Pragmatist thinkers may differ on the impact of the social and historical nature of knowledge, varying from those who follow in Peirce's footsteps by formulating a Pragmatic theory of Truth and Objectivity (S. Haack, 1998; Misak, 2013; Rescher, 2007) and those who following after James, most prominently Rorty (1982, 1999), consider such a project moot. Bacon (2012) summarizes Rorty's take on the role of truth indicating that "Rorty does not abandon the concept of truth. It is not a goal of inquiry because it is impossible for us to know that we have reached it, yet he insists upon the cautionary use of truth in order to contrast it with justification." Rorty argues simply that "*truth* is just the property that all true statements share" (Rorty, 1982), of which little of interest can be said. Nevertheless, despite these differences, they all share a rejection of the idea of a neutral and ahistorical foundation of knowledge.

As Brandom (2003), puts it, we are all engaged in *the game of giving and asking for reasons*, a point that can be traced back to Sellars's (1956) rejection of *the myth of the given*:

> The essential point is that in characterizing an episode or a state as that of *knowing*, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says. (p. 298-299)

Abandoning Representationalism

Another idea worth highlighting in the pragmatic tradition is the abandonment of representationalist theories, which are, in the words of Bernstein (2010), "theories that presuppose that the mind has impressions or ideas that represent objects that are 'outside of the mind' " (p. 140). Representational theories promote what Dewey (1929/1960) referred to as *the spectator theory of knowledge*; according to this theory, we human beings are observers of our environment, where we endeavor to collect information in order to develop ever more exact descriptions of nature. This focus on the accuracy of the representation has been challenged by Pragmatists, by proposing that we stop understanding theories and language as means for the accurate portrayal of nature, and start considering them in their role as instruments for manipulating nature (James, 1907/1995). In the words of Rorty (1999):

> Pragmatists insist on nonocular, nonrepresentational ways of describing sensory perception, thought and language, because they would like to break down the distinction between knowing things and using them. (p. 50)

The breakdown of the distinction between knowing and using that Rorty describes invites us to reconsider the criteria that we rely upon for judging the quality of our theories. In particular, it prompts us to abandon the idea that theories should be evaluated in terms of similarity to nature.[5] Pragmatist philosophers are not the only ones who advocate the move away from representationalism, and proponents of an alternative way of conceiving knowledge can also be found in cognitive sciences (Hutchins, 2000; Núñez & Freeman, 1999; Shanon, 1993; Varela, Thompson, & Rosch, 1991). An interesting example of a nonrepresentational perspective in cognitive psychology can be found in the work of Eleanor Rosch and her theories on concepts and categorization (Rosch, 1978, 1999; Rosch & Lloyd, 1978). As E. Margolis and Laurence (2014) indicate, it is widely accepted that concepts "are crucial

[5] From the pragmatic perspective I am proposing, your goals can imply empirical validation criteria, without having to make reference to a metaphysical reality. I can say, look, (a) historically the use of statistical fit has been shown an effective way of selecting among models, or, (b) the field of statistics, which has been shown to be extremely useful in a number of disciplines in the past, suggests this criterion for judging a model. These are two potential rationales for using it (there may be many others), without appeals to "reality."

to such psychological processes as categorization, inference, memory, learning, and decision-making"; however, the specifics of what concepts are, how they come to be formed, and how they fulfill their role in thought has become a controversial matter since the second half of the 20th century (Laurence & Margolis, 1999). Before the 1950's the perspective that predominated in the western tradition since Plato (see *Euthyphro*) was the *classical theory of concepts* or the *definition view*, according to which a conceptual category was a logical set clearly defined by a list of necessary and sufficient conditions (E. Margolis & Laurence, 2014). As Rosch indicates, a corollary of this perspective is that "all instances possessing the criteria attributes have a full and equal degree of membership" (Rosch & Mervis, 1975, p. 574). This classical/definitional perspective promotes thinking about concepts as representations of the outside world in the spirit of the spectator theory of knowledge. In her work, Rosch has proposed the *graded structure/prototype view*, a model in which:

> All categories show gradients of membership; that is, subjects easily, rapidly, and meaningfully rate how well a particular item fits their idea or image of the category to which the item belongs. Such judgments are the hallmark of the graded structure/prototype view.(Rosch, 1999, p. 66)

Rosch is a cognitive psychologist, not a Pragmatic philosopher, but just as with Wittgenstein, I think that her work displays a distinct Pragmatic affinity. Accordingly, I will use her theory of graded structure and prototypes as an additional instrument in my toolbox when outlining a Pragmatic perspective on measurement.

Theories as Tools

But if we abandon the spectator theory of knowledge. and we stop using similarity to reality as our criteria for evaluating our theories, then what? How can we assess the quality of our knowledge? According to Dewey (1920/2004), there is a clear alternative:

> As in the case of all tools, [theories'] value resides not in themselves but in their capacity to work shown in the consequences of their use. (p. 83)

To put this in idea in concrete terms, let us think of a specific tool; say, a hammer. Do you decide that a hammer is good because of its resemblance to a "true" hammer? Probably not. Most likely, you will judge the quality of a hammer in terms of its capacity for assisting you in the completion of a task, such as driving in a nail while assembling a shelf. Pragmatism emphasizes the idea that we can consider our theories in the same light, and that instead of focusing on how closely they resemble nature, we are better off by asking ourselves how useful they are to the accomplishment of our goals. From this perspective, theories, and language in general are considered to be "adaptation[s] to the environment, a set of tools for dealing with the causal pressures exerted by the environment, rather than a way of mirroring it" (Tartaglia, 2007, p. 214).

In sum, pragmatism invites us to evaluate our theories in terms of how they can help us cope with nature, as opposed to how accurate they are in representing it.

Pragmatism Recapped

This set of central ideas covered in the previous section, including the primacy of practice, fallibilism, anti-skepticism, the historical and social nature of knowledge, the anti-representational stance, and the conceptualization of our theories as tools, are nicely summarized by Menand (2001) in his book *The Metaphysical Club: A Story of Ideas in America*. In it, Menand discusses what he sees as the commonalities between James, Peirce, Dewey and Oliver Wendell Holmes, and I think captures the gist of the Pragmatic agenda by stating that:

> We can say that what these four thinkers had in common was not a group of ideas, but a single idea—an idea about ideas. They all believed that ideas are not 'out there' waiting to be discovered, but are tools—like forks and knives and microchips—that people devise to cope with the world in which they find themselves. They believed that ideas are produced not by individuals, but by groups of individuals—that ideas are social. They believed that ideas do not develop according to some inner logic of their own, but are entirely dependent, like germs, on their human carriers and the environment. And they believed that since ideas are provisional responses to particular and unreproducible circumstances,

> their survival depends not on their immutability but on their adaptability. (p. xi)

This is all fascinating, but, how does this relate to measurement at all? In the next section I discuss how I see these ideas contributing to our understanding of measurement.

2.3 What is Measurement?

There are indeed many, many definitions of measurement that have been proposed, influenced by the multiple areas in which the idea of measurement has been brought to bear. For a glimpse of the current widespread use of the concept of measurement we can look to the variety of areas of application of metrology, defined by the International Bureau of Weights and Measures as "the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology," which include manufacture of goods, international trade, satellite navigation, diagnosis in health related fields, and commerce (Bureau International des Poids et Mesures, 2014). A similarly broad scope is represented in the twenty-four technical committees of the International Measurement Confederation, which include areas such as the measurement of force, mass and torque (TC3), measurements in biology and medicine (TC13), and the measurement of human functions (TC18), the last one encompassing the "measurement, analysis and modeling of human functions, movement, perception and cognition" (International Measurement Confederation, 2014). Interest in measurement goes beyond the physical sciences, playing a prominent role in economics (Boumans, 2007; Spengler, 1961), and education (Brennan, 2006; Lindquist & Thorndike, 1951), as well as sociology and psychology (Boring, 1961; Duncan, 1984; Lazarsfeld, 1961), leading to the formation of organizations such as the Psychometric Society concerned with "the advancement of quantitative measurement practices in psychology, education and the social sciences" (Psychometric Society, 2014).

Historically, a few scholars in mathematics, physics, philosophy and psychology have striven to present a unified definition of measurement. Some, like Michell (1997b), have turned to the history of the concept in the physical sciences to "properly

define" what "scientific measurement" is. Others, like Berka (1983), for instance, have embraced the multiplicity of applications and have sought to distill the *essence* of measurement through the examination across context of applications:

> It is much more important for the theory of measurement if we expound what can be meaningfully said about measurement from various standpoints: if we state what are the general and specific characteristics helping us to grasp the core of the method, what its role is in the process of scientific knowledge, under what conditions measurement can be legitimately applied, *and what indeed measurement is objectively* [emphasis added]. (p. 14)

Regardless of their strategy, the motivations for establishing (or discovering) *what measurement is* include attempts to identify solid foundations for the methods of science, to establish a demarcation criterion for what could and could not be measured, and to establish an encompassing definition that would facilitate communication across diverse knowledge domains.

Throughout history, the technical understanding of measurement has been influenced by key philosophical perspectives such as the "classical view" of measurement (Michell, 2005), operationalism (Bridgman, 1927), the representational theory of measurement (Krantz et al., 1971/2007; Pfanzagl, 1968), as well as more methodological perspectives arising from the practice of researchers in the social sciences, such as latent variable perspective (Lazarsfeld & Henry, 1968; Lord & Novick, 1968; Rasch, 1960/1980), and physical sciences and engineering, represented by metrology. The next sections present a brief overview of these influential strands, but should not be in any way construed as an exhaustive review of the myriad of perspectives that have been proposed about the concept of measurement.

2.3.1 Classical Theory of Measurement

The Classical Theory of Measurement (CTM[6]) holds, generally speaking, that objects in reality have properties and that some of those properties are quantities. Mea-

[6]Not to be confused with the Classical Test Theory in psychometrics, despite some overlap in terminology.

surement is seen as an activity concerned with the discovery and study of those quantitative properties.

Joel Michell (1997b, 1999, 2005), one of the strongest proponents of the return to the definitions of measurement laid out in the CTM, relies heavily on meticulous historical analysis of the concept of measurement and its origins in order to establish clear demarcation criteria regarding the correct way to practice measurement.

In his treatment of the CTM, Michell (2005) is explicitly committed to realism, which he defines in terms of commitment to a theory of truth that specifies "that a proposition is true if and only if things are as proposed" (p. 286). About this philosophical foundation, Michell states:

> This literal concept of truth commits us to the existence of things... existing in time and space, independently of observation. The conviction that there is an objective, spatiotemporally structured world is *metaphysical realism* and the implication that things are logically independent of observation is *epistemological realism*. (p. 286)

From this perspective, Michell adopts a set of distinctions that go back to Aristotle's philosophical realism, chief among them the difference between *quantity* and *quality* as two of the fundamental categories that describe all that exists. Aristotle writes that "it is the distinctive mark of quantity that it can be called equal and unequal" (p. 17, trans. 2001) and "the fact that likeness and unlikeness can be predicated with reference to quality only, gives to that category its distinctive feature" (p. 27, trans. 2001). Aristotle further differentiates between two kinds of quantities, *multitudes* or *pluralities* which are discrete quantities[7], and *magnitudes* which are continuous quantities, stating that "a quantum is a plurality if it is numerable, a magnitude if it is measurable" (p. 27, trans. 2001).

In the context of these classical distinctions, Michell (1990, 2005) traces the origin of CTM back to Book V of Euclid's *Elements* (trans. 1956). The first three definitions in Book V (p. 113–14) where a measure is defined as a ratio of one magnitude to another are:

[7]According to Michell (2011) the analysis of multitudes through frequencies is indeed quantitative, but not measurement: "Counting is a quantitative method, but not measurement (although it may be involved in processes of measurement)..." (p. 255). Footnote 19 (page 57) briefly touches on how is counting considered under the Pragmatic perspective in this paper

1. A magnitude is a *part* of a magnitude, the less of the greater, when it measures the greater.
2. The greater is a *multiple* of the less when it is measured by the less.
3. A *ratio* is a sort of relation in respect of size between two magnitudes of the same kind.

The foundations of the CTM laid by Euclid in Book V and Book VII of the *Elements* were (much) latter fully axiomatized at beginning of the 20th century by Otto Hölder (Michell & Ernst, 1996, 1997)[8], and one year later an alternative axiomatization was proposed by Huntington (1902). Rooted in this treatment of magnitude, ratio and measure by Euclid and the axiomatization by Hölder, Michell (1997b) defines measurement as "the numerical estimation of the ratio of a magnitude of a quantitative attribute to a unit of the same attribute" (p. 383), going on to call the task of identifying whether a property of interest is or not a quantity the *scientific task* of measurement, as opposed to the *instrumental task* which focuses on the practicalities of conducting measurements.

The CTM establishes then a clear demarcation criterion for what constitutes measurement, which, according to Michell (1997b, 1999, 2005) had historically characterized measurement in the physical sciences by virtue of the discovery of quantities and the development of the appropriate instrumentation to determine their ratios.

According to Michell (2012a), all attributes can be classified into one of these three "structural possibilities":

> (1) *classificatory attributes* (with heterogeneous differences between categories);
> (2) *heterogeneous orders* (with heterogeneous differences between degrees); and
> (3) *quantitative attributes* (with thoroughly homogeneous differences between magnitudes). (p. 1)

Under CTM, the existence of quantity is the *sine qua non* of measurement. If the properties under study are not in themselves quantitative (i.e. if the answer to the

[8]Hölder's original work was published in 1901 but was not translated into English until the late 1990's by Joel Michell and Catherine Ernst.

scientific task of measurement is that they are classificatory or heterogeneous orders), there is no amount of work on the instrumental side that could ever justify the use of the word "measurement" in the study of such a property. It follows from this that if the true nature of a psychological property is qualitative, or ordinal for that matter, measurement of such a property is by definition impossible (Figure 2.1).

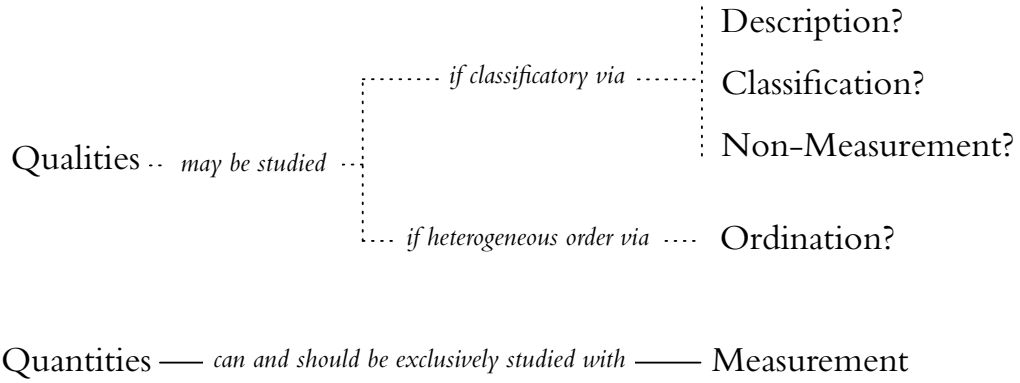


Figure 2.1: Measurement and Attributes in the Classical Theory of Measurement

The concept of measurement, Michell (1997b, 1999, 2005) claims, was then co-opted by the emergent discipline of psychology in the second half of the 19th century, and replaced by alternative, false definitions of measurement, that exempted psychology from seriously considering and conducting the scientific task of measurement, and casting doubt about the adequacy of any development on the instrumental task:

> *There is no evidence that the attributes that psychometricians aspire to measure (such as abilities, attitudes and personality traits) are quantitative.* Mainstream psychologists presumed that they are and that psychometric tests are suitable to measure them. All the evidence is that these attributes are merely ordinal, and while psychologists are free to speculate about whether they might be quantitative, the fact is that, after a century, such speculations remain empty. (Michell, 2011, p. 245)

Of course, the position that psychology, and potentially other social sciences, cannot engage in measurement because the attributes that are purported to be measured are not actually quantitative has a long history, and is known as *the quantity objection* (Boring, 1921):

> Introspection, the objection runs, does not show that a sensation of great magnitude ever contains other sensations of lesser magnitude in the way that a heavy weight may [supposedly] be made up of a number of smaller weights. "Our feeling of pink", said James, "is surely not a portion of our feeling of scarlet; nor does the light of an electric arc seem to contain that of a tallow-candle in itself." "This sensation of 'gray'," remarked Külpe, "is not two or three of that other sensation of 'gray' " (p. 453).

According to this objection, psychological properties are qualitative or in some cases ordinal, and therefore not subject to measurement; as H. M. Johnson (1936), a psychologist who adhered to this perspective asserts: "There are questions concerning human traits which are interesting; they may be important; they may be answerable; but they probably will remain unanswered as long as they are attacked by attempts to measure what is *intrinsically non-measurable* [emphasis added]" (p. 351).

However, all these criticisms to the core idea that psychological properties could be measured caused little impact and were ignored by a large majority of practitioners and researchers throughout the last century and a half of psychology (Cliff, 1992; Hornstein, 1988; Michell, 1999).

2.3.2 Operationalism

Operationalism (Bridgman, 1927) is perhaps the philosophical and methodological approach that presents the starkest contrast to the CTM. Developed originally by physicist P.W. Bridgman, this approach originated as his attempt to avoid the theoretical problems in physics that had been brought to light with the introduction of Einstein's theory of special relativity, of which Bridgman (1927) states:

> It was a great shock to discover that classical concepts, accepted unquestioningly, were inadequate to meet the actual situation, and the shock of this discovery has resulted in a critical attitude toward our whole conceptual structure which must at least in part be permanent. (p. 1)

Bridgman's efforts on this issue were focused on the elimination of ambiguity and confusion of concepts in physics due to the ever-present risk of extrapolation beyond

areas directly examined by our experience. In his introduction of operationalism in *The Logic of Modern Science* (1927), he illustrates these issues by discussing the case of length, which he considers to be defined by the operations used to measure it.[9] To achieve his goal, Bridgman introduced the doctrine of *operational analysis*, according to which "we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations" (Bridgman, 1927, p. 5).

In philosophy of science, operationalism has been extensively criticized on a range of issues, ranging from challenges of attempting to equate the meaning of a theoretical term to a set of operations to more basic issues such as the lack of clarity of what should count as an operation (Chang, 2009). However, for the purpose of this paper, the most glaring issue with a strict operationalist perspective is the inability to coherently formulate the idea of multiple measures of the same attribute (Markus & Borsboom, 2013), leading us to conclude that length, as measured by a ruler in centimeters, is an entirely different concept from length as measured in terms of parsecs by a telescope. This position and its consequences for measurement were explicitly advocated by Dingle (1950) in his paper *A Theory of Measurement*, where he stated:

> We must return to simplicity and describe what we do as faithfully as possible, without introducing conceptions which have outlived their usefulness. I therefore suggest the following definition: *A measurement is any precisely specified operation that yields a number.*

To be fair, both Bridgman (1927) and Dingle (1950) addressed the possibility of equating different sets of operations or attributes when their results could be compared, but considered this as a matter of practical convenience. Bridgman in particular still cautioned against assuming that such an equating procedure would amount to conceptual identity, an especially tempting possibility once we have summarized the treatment of a nanometer and a parsec as the same variable in a simple equation:

> The equations of motion make no distinction between the motion of a star into our galaxy from external space, and the motion of an electron

[9] "The concept of length is therefore fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as and nothing more than the set of operations by which length is determined" (Bridgman, 1927, p. 5).

> about the nucleus, although physically the meaning in terms of operations of the quantities in the equations is entirely different in the two cases.... What we would like is some development of mathematics by which the equations could be made to cease to have meaning outside the range of numerical magnitude in which the physical concepts themselves have meaning. (Bridgman, 1927, p. 63)

Despite the problems that operationalism encountered as a philosophical doctrine, it was warmly adopted, and adapted, within psychology (Grace, 2001; Green, 1992). The influence of operationalism is still felt today through the idea of an *operational definition* of which Koch (1992) has acerbically remarked:

> The notion, as it now functions in psychology, is part of an empyrean of jargon that instructors draw upon in order to give students the illusion that they are studying a *science*. (p. 262)

As we have discussed, a strict operationalist approach eliminates the separation between defining a concept and measuring it, by making the former synonymous with the latter, and in principle raises the possibility of measuring any property, regardless of its structure (Figure 2.2).

Measurement ——— *defines and can be used for studying* ——— Attributes

Figure 2.2: Measurement and Attributes in Operationalism

Although *prima facie* this may seem an invitation for an *anything goes* approach to measurement (and theory-building for that matter), it is important to remember that Bridgman, a renowned and Nobel winning physicist, was formulating his doctrine within the nomological network of physics and their shared practices. As Koch (1992) indicates, Bridgman presented his theory in this context as a way to help clarify concepts that already were part of the theoretical repertoire of physics, not as a way to create new concepts by defining them in an operational manner. Moreover, in *The Logic of Modern Science* (1927) Bridgman presents his arguments from a conception of measurement rooted in the tradition of physical measurement based on empirical operations. Alas, when operationalism was introduced to psychology, it would jettison

some of these background assumptions and would be paired with a radically different approach to measurement.

2.3.3 The Representational Theory of Measurement

A third major strand that has influenced measurement theory, and arguably the predominant theory at the moment (Michell, 1993; Narens, 2013), at least among philosophers of measurement in the social sciences, is the Representational Theory of Measurement (RTM). This approach can be traced back originally to the work of Russell on the relation of number and quantity (1897), where he argued the independence of the notion of *number* from the notion of *quantity*, as opposed to the traditional view where numbers were based on quantity. In his review of Russell's 1903 critique to the traditional account of the number-quantity relation, Joel Michell (1997a) summarizes the central departure in the following way:

> The fundamental issue is whether or not ratios of magnitudes should be regarded as *instances* (Armstrong, 1987) of real numbers. If they are understood in this way then the logic of the application of arithmetic to reality is that of *instantiation* rather than that of *representation*. (p. 270)

Russell (1903), based on this divorce between the concept of number and quantity, went on to state that "measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real, as the case may be" (p. 176), laying the groundwork for the representational perspective that would be later popularized by the physicist Norman Campbell (1920), and that achieved its maturity with the work of Pfanzagl (1968) and the publication of the *Foundations of Measurement* volume by Krantz et al. (1971/2007).[10]

In contrast to the CTM, which is centered around the idea of estimating ratios of magnitudes, the RTM focuses on the "correlation with numbers of entities which are not numbers" (Nagel, 1931, p. 313). This shift implies, in the words of Michell (1993), that "[the] numbers used in measurement are contributed to the measurement

[10]See Michell (1999) for an overview of the theoretical origins of the RTM and its departure from the classical conception of measurement in the work of Russell, Campbell and Nagel.

situation, rather than discovered within it" (p. 189). In practical terms, one of the main consequences of the decoupling of ratio estimation from measurement in the RTM was the "liberalization" of the concept of measurement to include ordering and classification, as was advocated in the version of the RTM developed by S.S. Stevens (1946).

Campbell: Between the Classical and Representational Theories

Campbell (1920) dedicated the second part of his book *Physics: The Elements* to the topic of measurement, beginning with the following definition: "Measurement is the assignment of numerals to represent properties" (p. 267)[11]. In his theory of measurement, Campbell indicates that although an object may possess many attributes or qualities, not all of them can be measured. Campbell proposed that what made a property or quality measurable was our ability to relate that property to numbers in a very specific manner, namely, by empirically fulfilling relations of *order and additivity*.

Campbell emphasized that an empirical process of concatenation was necessary in order to test experimentally whether the property conformed to the mathematical laws of order and addition. Based on this empirical criterion, he declared that "the difference between properties that are and those that are not capable of satisfactory addition is roughly that between quantities and qualities" (p. 267).

On this basis, Campbell then drew the distinction between *fundamental measurement* and *derived measurement*. Fundamental measurement applied to properties for which an empirical concatenation operation was available, with length, mass and volume as examples. Derived measurement encompassed properties that lacked an empirical concatenation operation, but could be measured via other fundamentally measured properties. Density, for example, can be measured through mass and volume.

The focus on the need for empirical concatenation operations made Campbell's theory ideally suited to the work of physicists, but presented a serious challenge for social science disciplines such as psychology.

[11] Berka (1983) points out that this is not Campbell's only definition throughout his work, characterizing it also as "the process of assigning numbers to represent qualities" (Campbell, 1920, p. 267), "the assignment of numerals to represent properties according to scientific laws" (Campbell, 1928, p. 1), and "the assignment of numerals to things so as to represent facts or conventions about them" (Campbell, 1940, p. 340).

Campbell's theory can be considered a mixture between the CTM and RTM (Michell, 1993, 1999), which on the one hand embraced classical intuitions such as the idea that only quantities could be subject to measurement, while on the other adopting the idea of measurement as correlation between empirical properties and a numerical system. It was on this aspect of the theory that some psychologists would rely to answer the challenge presented by Campbell's theory. On this unwitting outcome of Campbell's work Berka (1983) indicates:

> Although Campbell's monograph was programmatically opposed to any further extension of measurement, it nevertheless became the foundation of a general theory of measurement and the stimulation for its advancement of theoretical and methodological conceptions outside physics. (p. 13)

Stevens: Between the Representational and Operationalist Theories

S.S. Stevens' theory of measurement was developed as a direct answer to Campbell and his influential theory of fundamental and derived measurement. Stevens' response was presented as a challenge to the conclusions presented by a committee of the British Association for the Advancement of Science (BAAS), which had been tasked in 1932 with answering the question of whether sensations could be measured, a task that was conducted under the influential definitions of measurement proposed by Campbell, which demanded the existence of empirical concatenation operations.

In other words, the BAAS committee had been asked to judge whether measurement was possible in psychophysics. This judgment was conducted under Campbell's influential theory of measurement, which was tailored after the physical sciences.

After working for the better part of a decade, the BAAS committee reported its results, which were described in the journal *Nature* as follows:

> It now seems that agreement is impossible. This is scarcely surprising when there is disagreement as to the meaning of "measurement". If it is postulated that this term must be limited to its applicability in physics, then this would rule out the use of the word in relation to much psychological work. Two extreme views hold the field, and a close examination

> of these views by a member who holds an intermediate view leads to the conclusion that they cannot be reconciled. (Nature Publishing Group, 1939, p. 973)

In his 1946 paper *On the Theory of Scales of Measurement*, Stevens directly challenged the idea that the definition of measurement depended on the existence of empirical concatenation, indicating that "perhaps agreement can better be achieved if we recognize that measurement exists in a variety of forms and that scales of measurement fall into certain definite classes" (p. 677). Stevens went on to present this "variety of forms" by fully embracing the representational spirit of Campbell's definition while jettisoning its connection to the physical sciences, stating what by now has become a commonplace definition of measurement in psychology:

> We may say that measurement, in the broadest sense, is defined as the assignments of numerals to objects or events according to rules (p. 677).

While keeping the representational notion of the assignment of numerals to objects or events, Stevens switched the focus from the *empirical testing of the mapping* between properties and numbers, Campbell's original emphasis, to *the rules of that govern the mapping*: "The problem as to what is and is not measurement then reduces to the simple question: What are the rules, if any, under which numerals are assigned?" (p. 680). This focus on the mathematical properties that the rules of assignment entailed is at the basis of his taxonomy of *scale types*: nominal, ordinal, interval and ratio scales (Stevens, 1946).[12]

Each one of these four types of scales was defined according to Stevens by the kind of mathematical group structure associated with it and the set of permissible statistics that the scale would support (Stevens, 1946). Table 2.1 reproduces the table originally presented in Stevens (1946).

Stevens (1946) taxonomy includes forms of non-quantitative measurement (i.e. the nominal and ordinal scales), in addition to the interval and ratio quantitative scales,

[12]It is worth noting that in his 1946 paper, Stevens discusses and details basic empirical operations associated with each one of the scale types. However, his summary regarding the characterization of measurement focuses exclusively in the rules of assignment, and ignores completely any empirical requirements.

Table 2.1: Scale Types as Presented in Stevens (1946)

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution.	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater and less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function.	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

but he barely acknowledged this considerable expansion of the term measurement beyond quantity; In Stevens' treatment of measurement, the concept of quantity is relegated to the background[13].

While Campbell's definition retained the idea that measurement could only be conducted when dealing with certain properties and not others, and at least implicitly indicating that this difference was somehow dependent on a property of nature itself, Stevens' new definition severed this connection, focusing solely on the procedure used to generate the assignment, allowing measurement of any property as long as a mapping was established according to rules.

Accordingly, Campbell's clear distinction between what is measurable and what is not measurable, with the quantitative on the former side, and the qualitative and ordinal on the latter, was also disregarded, effectively declaring that any kind of property was potentially measurable under one of the different scale types if the researcher developed a rule-based procedure to deal with it (Figure 2.4). In other words, measurement was achieved by the sheer fact of establishing a rule-governed measurement

[13]Stevens (1946) states at the beginning of his discussion on interval scales (after discussing both nominal and ordinal scales) that "with the interval scale we come to a form that is "quantitative" in the ordinary sense of the word." This is one of the only three allusions to quantity in his paper, and the only one outside direct quotes to the BAAS report.

procedure. If this sounds familiar, it is because it echoes Bridgman's operationalist doctrine described in section 2.3.2 (Michell, 1990).

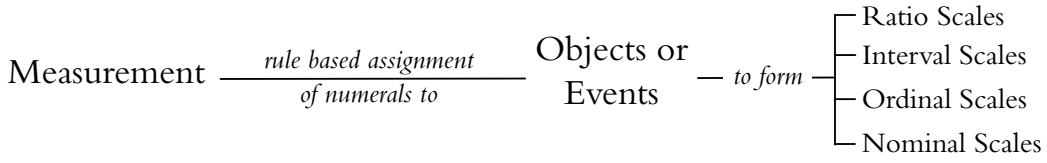


Figure 2.3: Measurement and Attributes in Steven's Theory of Measurement Scales

It is worth pointing out, as Hand (2004) does, that even though Stevens' is usually cited (or criticized) for his 1946 definition of measurement "as the assignments of numerals to objects or events according to rules" (p. 677) and its sole focus on the rules of assignment, this was hardly the only aspect of the definition of measurement that he wrote about. In the very same 1946 article, Stevens highlights one of the definitions proposed by Campbell (1920):

> To the British Committee, then, we may venture to suggest by way of conclusion that the most liberal and useful definition of measurement is, as one of its members advised [i.e. Campbell], "the assignment of numerals to things so as to represent facts and conventions about them." (p. 680)

In contrast with the sole focus on the rules of assignment, this reference to Campbell's definition brings to bear the importance of "representing facts and conventions," a criterion that goes beyond the use of a non-random rule of assignment. Stevens' would continue to emphasize the "representation of nature" stating for instance in 1951 that "the type of scale achieved when we deputize the numerals to serve as representatives for a state of affairs in nature depends upon the character of the basic empirical operations performed on nature" (p. 23). However, this disconnect between his more radically liberal statements about measurement and his more nuanced emphasis in the preservation of properties in nature through mathematical representations is unsurprising given Stevens' penchant for extremely memorable one-liners, such as "*in its broadest sense, measurement is the business of pinning numbers on things*" (Stevens,

1958), a statement which he immediately followed with the much less remembered qualification:

> This process turns out to be a fruitful enterprise only because some degree of isomorphism obtains between the empirical relations among the properties of objects or events, on the one hand, and some of the properties of the number system on the other.

In any case, it was Stevens' emphasis on rules of assignment in his 1946 definition and its radical reinterpretation of the concept of measurement that would be remembered, encompassing many psychology methodologies under a new, much more liberal, theory of scales of measurement. However, it is important to note that Stevens' work presents a point of departure for two distinct traditions regarding the development and application of this new representational understanding of measurement, one focusing on his definition based on the rules of assignment, the other based on his stipulation on the need for isomorphism with empirical relations.

In the first tradition, Stevens' theory of scales types and its mixture between representational and operational measurement theory was adopted largely unchanged, and was further popularized to the point of becoming a sort of received view of measurement in psychology and the social sciences in general; Duncan (1984, p. 125-126) stated on the ubiquity of this definition that "after all, most of what most of us know about measurement theory we learned from Stevens, if not directly then via some textbook author's secondary presentation." This strand embraced its operational nature (even though its philosophical background was not always explicitly acknowledged) (Green, 1992, 2001). Torgerson (1958, p. 22) colorfully named this approach "measurement by fiat," but despite the skepticism that the name might invite, Torgerson did not dismiss this approach, stating that:

> It has led to a great many results of both practical and theoretical importance. For example, a major share of the results of the field of mental testing and of the quantitative assessment of personality traits has depended upon measurement by fiat. Measurement of morale, efficiency, drives, and emotion, as well as most sociological and economic indices, is largely measurement of this type (Torgerson, 1958, p. 24).

However, not everyone was persuaded of the benefits of adopting this characterization of measurement. For example, it was later mocked by Wolins (1978, p. 1-2) as based simply in the argument that "we need interval measurement for certain types of statistical analysis, we don't really have it but we will pretend we do and everything will be all right." Stevens' theory of scale types would continue to exert a large influence, despite its critics in psychology (cf. Michell, 1999), sociology (cf. Duncan, 1984), and social statistics methodology (cf. Velleman & Wilkinson, 1993), and continues to be the default definition of measurement even today in psychological methods textbooks (see for instance Elmes, Kantowitz, & Roediger, 2012; Nestor & Schutt, 2014).

In a separate, second tradition, the representational theory continued to evolve, taking Stevens' theory of scales as a starting point, and eventually leading to the formulation of axiomatic measurement theory (Krantz et al., 1971/2007; Luce, Krantz, Suppes, & Tversky, 1990/2007; Pfanzagl, 1968). This more mature rendition of the representational approach would emerge as an alternative to the operationalist strand by embracing that "the more fruitful view is that the nature of a scale is defined by the relations exhibited among empirical variables" (Cliff, 1992, p. 186), distancing itself from the operationalist commitments of the first strand.

Axiomatic Measurement Theory

An extensive mathematical basis for representational measurement was developed in the form of Axiomatic or Abstract Measurement Theory (AMT; Krantz et al., 1971/2007; Luce & Tukey, 1964; Scott & Suppes, 1958; Suppes & Zinnes, 1963)[14].

According to Narens and Luce (1986, p. 173), AMT holds at its core four conditions that are necessary for measurement: (1) an "ordered relational structure $\mathscr{X} = \langle X, \succsim, S_1, \ldots, S_n \rangle$, where $\succsim, S_1, \ldots, S_n$ are the *primitives* of the structure", (2) a set of axioms that describe the empirical structure (i.e empirical laws), (3) a "a numerically based relational structure $\mathscr{R} = \langle R, \geq, R_1, \ldots, R_n \rangle$, where R is a subset of the real numbers and the R_i are relations and operations of comparable types to the corresponding empirical ones", and the fourth and last one:

[14] A detailed overview of AMT is outside the scope of this paper. Introductory texts to the basics of the theory can be found in Michell (1990), Narens and Luce (1986), and Narens (2013). An extensive treatment can be found in the three volumes of Foundations of Measurement (Krantz et al., 1971/2007; Luce et al., 1990/2007; Suppes, Krantz, Luce, & Tversky, 1989/2007)

> Which accomplishes measurement, is the proof of the existence of a structure preserving mapping from $\mathscr{X}$ into $\mathscr{R}$. We refer to $\mathscr{X}$ as the *empirical* or *qualitative structure*, $\mathscr{R}$ as the *representing structure*, and the structure-preserving mapping as a *homomorphism* or a *representation*. The collection of all homomorphisms into the same representing structure is referred to as a *scale*. (p. 173)

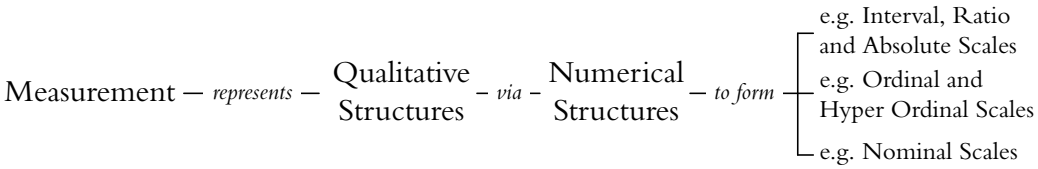


Figure 2.4: Measurement and Attributes in Representational Measurement Theory

This framework sidestepped any debate about quantitative vs qualitative properties that concerned Campbell (and adherents to the CTM), by defining all measurement as based on the mapping of the *qualitative empirical structure* into a *representing numerical structure*. Instead of focusing on the "kinds of properties", the AMT followed Stevens' lead in using the concept of scales, now a formalized concept, developing it beyond Stevens (1946) original taxonomy.

Under this new formalization of the concept of measurement, AMT subsumed the focus on physical variables of Campbell's fundamental measurement theory, replacing it with a more general redefinition idea of "fundamental measurement" based on numerical assignment, jettisoning Campbell's original requirement of an empirical concatenation operation as the sole foundation of measurement (Scott & Suppes, 1958; Suppes & Zinnes, 1963). This reframed fundamental measurement as a part of a larger, psychology friendly, measurement theory (Narens & Luce, 1986), of which Schönemann (1994) has commented:

> ...Campbell's version became a special case which the authors [Suppes and Zinnes] called "extensive measurement". This change in semantics enabled psychologists to charge Campbell with having been fundamentally mistaken: of course social scientists have fundamental measurement, they just do not have any extensive measurement.

A key achievement of this expansion of the concept of fundamental measurement was the development of Conjoint Measurement (Luce & Tukey, 1964), which showed how to establish an interval scale based on the relations between a set of ordinal variables, without using an empirical concatenation operation (Narens & Luce, 1986). In other words, AMT presented a mathematical foundation for using a quantitative scale solely on the basis of having identified a set of ordered variables and a specific set of relations. In principle, this was a colossal breakthrough, offering finally an answer to the challenge presented by Campbell and his theory to the possibility of measurement in psychology, and providing an overarching measurement framework across the physical and social sciences.

Nevertheless, the idea of psychological measurement that became commonplace was based on Stevens' original definition (i.e. measurement by fiat[15]), largely disregarding the advances made by AMT. In any case, the lack of impact of AMT on the practice of psychological measurement was memorably labeled "the revolution that never happened" by Cliff (1992), where he lamented that while in psychology "abstract measurement theory has to be regarded as one of its major intellectual achievements, viewed in terms of the power of thought that was required to achieve it and as a key pillar of the philosophy of science.... *it is clear that the influence of abstract measurement theory has been small* [emphasis added]" (p. 186). Although AMT has been adopted to conduct research on some areas of psychology (Narens & Luce, 1993), its promise to serve as a foundation for measurement across the social sciences has largely been unfulfilled. However, it is worth noting that while advocates of the AMT tend to frame this issues as a problem of the larger psychological (and social science) community ignoring AMT's accomplishments, there is a case to be made that it was AMT that was developed while ignoring large parts of the measurements conducted in the social sciences and even some physical measurements (E. W. Adams, 1979), and that the sparseness in its adoption is a product of this "strategy":

> One reason for the relative paucity of connections between measurement theory and substantive theory in psychology may arise from the

[15] As an example of the extent to which AMT is considered by some scholars to be *the* definition of measurement, Dawes and Smith (1985) implicitly assumes that anything other than AMT is effectively measurement by fiat, which they label as *non-representational measurement*.

> fact that models for measurement have largely been developed independently as a body of abstract formal theory with empirical interpretations being left to a later stage. *The difficulty with this approach is that the later stage often fails to materialize.* [emphasis added] (Estes, 1975, p. 273)

From this perspective, the lack of adoption of AMT is less surprising, and can be attributed to some extent to the methodological and theoretical characteristics of the approach (cf. Schönemann, 1994; Schwager, 1991), such that:

> It is absurd to hold that these axioms state conditions which must be satisfied in order that measurement be possible, or that it be justified, since measurement clearly is possible and justified even though some of the measurement axioms are false. (E. W. Adams, 1966, p. 131)

In trying to understand "Why did the revolution not occur?" (p. 188), Cliff (1992) entertains several hypothesis, both related to the larger research community and to the AMT itself, to account for the lack of influence of AMT in psychology (and the social science at large for that matter). These include the lack of the necessary level of mathematical knowledge among researchers, the lack of connections between the abstract mathematic treatment and the empirical challenges confronted by researchers, the lack of examples that attest to the "empirical power" of AMT (see also the 1994 paper by Schönemann for an elaboration on these criticisms). But perhaps more importantly, Cliff touches on what he calls "The Error Problem":

> Another kind of bridging element that has been lacking is ways to deal with failure to conform to the axioms. Measurement theory says that if certain conditions hold, then scales of a given kind are defined. If not, they are not. *But data always contradict one or the other axiom*. What now? Error is always present in observation, yet how is it to be dealt with? How is one to deal with occasional failures of the cancellation axiom in conjoint measurement or transitivity in paired comparisons? The levels of variables are never infinitely fine, as is often required in the proofs. (p. 189)

And while Narens and Luce (1993) in general reject the charge that AMT has failed to produce practical applications that demonstrate it "empirical power," they acknowledge that, as opposed to statistics, "AMT focuses almost exclusively on structure, largely ignoring randomness" (p. 129), and that:

> Neither [AMT nor statistics] has any very clear idea about how to formulate the concept of randomness in a nonnumerical way. In particular, we simply do not know how to talk about randomness at a level involving only qualitative ordinal observations together with other qualitative relationships, such as combining two stimuli to form a third. (p. 129)

This issue had been previously raised in a more general manner by E. W. Adams (1979, p. 220): "A general difficulty with the current accounts is that they do not deal systematically with *error* and accuracy, which is in turn linked to the fact that they give no consideration to *variability* of objects over time, and the special role which *standards* play in measurement."

Although some efforts have been made to develop extensions to AMT that deal both with issues of accuracy (i.e. measurement with imperfect discrimination) through the use of semiorders (E. W. Adams, 1965; Krantz, 1967; Luce, 1956) and accounts for error through random variability (Domingue, 2014; Falmagne, 1976, 1980; Perline, Wright, & Wainer, 1979), this remains an issue to this day. Interestingly, the problem of dealing with error is related to another factor that Cliff (1992) believes contributed to the lack of the influence of AMT, namely, the "distracting development" (p. 189) of covariance structure analysis and latent variable models in general, which do not see randomness as a problem, but are—for good or ill—dependent on it.

2.3.4 Latent Variable Modeling

Latent Variable Modeling can be considered more as a methodological tradition than an organized measurement theory. Nevertheless, its advocates and practitioners played an important role during the 20th century in the popularization of statistical methods as a way of conducting measurement of psychological properties. The core intuition behind this tradition can be traced back to Charles Spearman, who

discussed how the correlation between multiple observed variables was not only interesting in itself, but allowed for the possibility that "another—theoretically far more valuable—property may conceivably attach to one among the possible systems of values expressing the correlation; this is, that a measure might be afforded of the *hidden underlying cause of the variations*" (1904b, p.74). Spearman relied on this insight to introduce what became known as the *common factor model*, applying it to the measurement of intelligence (Spearman, 1904a), adding the concept of a latent variable to psychology's methodological armamentarium and kickstarting the factorialist tradition of intelligence measurement.

At its core, the concept of latent variable modeling remains faithful to Spearman's initial use, namely, that we can reduce the complexity in a set of observed variables if we postulate a different, unobserved, underlying variable. This core idea behind latent variables has more recently been expressed under a general statistical framework as a "random variable whose realizations are hidden from us as opposed to manifest variables where these realizations are observed" (Skrondal & Rabe-Hesketh, 2004).

Latent variables as a statistical tool can have multiple interpretations, but in psychology, and the social sciences more generally, they have been recurrently used as a measurement methodology that enabled researchers to study a wide variety of constructs (Skrondal & Rabe-Hesketh, 2004). *Constructs* in this context is the name that is used to describe the attribute that the researcher is trying to measure (Messick, 1989; Wilson, 2005).

Although, it is important to point out that there is no official definition of measurement that has been endorsed, for instance, by mayor professional associations—as we will see is the case in Metrology—measurement in latent variable modeling is more often than not discussed in terms of Stevens' 1946 definition, and in some cases extending it. For instance, Lord and Novick (1968, p. 17) define measurement as "a procedure for the assignment of numbers (scores, measurements) to specified properties of experimental units in such a way to characterize and preserve specified relationships in the behavioral domain." Similar definitions, embracing the idea of measurement as *assignment* can be found, when measurement is defined, through the literature, for instance, Weitzenhoffer (1951), Guilford (1954, who attributes the definition to Campbell),Torgerson (1958), Crocker and Algina (1986), McDonald (1999), and de Ayala (2009). In this context, latent variable models are seen as a method of conducting this numerical assignment.

This approach to measurement has been described by van der Linden (1994) in the following manner:

> Modern measurement theory shows that we can go one step further and verify laws that explain observable data using only unmeasured variables. If these laws—or models, as modern measurement theory prefers to call them—are quantitative and empirically verified, then the unmeasured or latent variables have quantitative scales on which, as a byproduct, the positions of the objects are known. As the model contains only latent variables, measurement of them is not derived from other fundamentally measured variables—all variables are measured jointly, in relation to one another (p. 12)

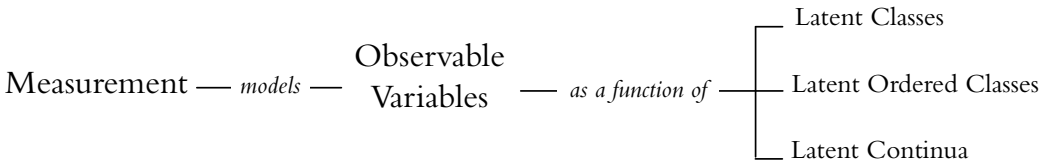


Figure 2.5: Measurement and Attributes in Latent Variable Modeling

Since its introduction in 1904, the study of latent variables has flourished, and expanded beyond the single, continuous latent trait methodology initially proposed by Spearman when modeling intelligence. Multiple frameworks have been proposed to account for the myriad of variations and areas of development (see Langeheine and Rost (1988), Heinen (1996), and Skrondal and Rabe-Hesketh (2004) for reviews of the extensive family of latent variable models). One important and relatively early framework, Latent Structure Analysis (Lazarsfeld & Henry, 1968, henceforth LSA), included continuous (latent trait models), categorical (latent classes), and ordinal (ordered latent classes) variables, illustrating how this methodology could be applied to model attributes of different structures (Figure 2.5).

An example of the treatment of latent variable modeling as a measurement approach is presented by de Ayala (2009) where he indicates:

> The measurement process involve deciding whether our latent variable, anxiety, should be conceptualized as categorical, continuous, or

> both. In the categorical case we would classify individuals into qualitatively different latent groups so that, for example, one group may be interpreted as representing individuals with incapacitating anxiety and another group representing individuals without anxiety. In this conceptualization the persons differ from one another in *kind* on the latent variable. Typically, these latent categories are referred to as *latent classes*. Alternatively; anxiety could be conceptualized as continuous. From this perspective, individuals differ from one another in their *quantity* of the latent variable. Thus, we might label the ends of the latent continuum as. say, "high anxiety" and "low anxiety." (p. 2)

Many different structures of the latent variable can be hypothesized, and models have been proposed to deal with each one of these structures. For instance, see five the different possibilities presented in Figure 2.6.

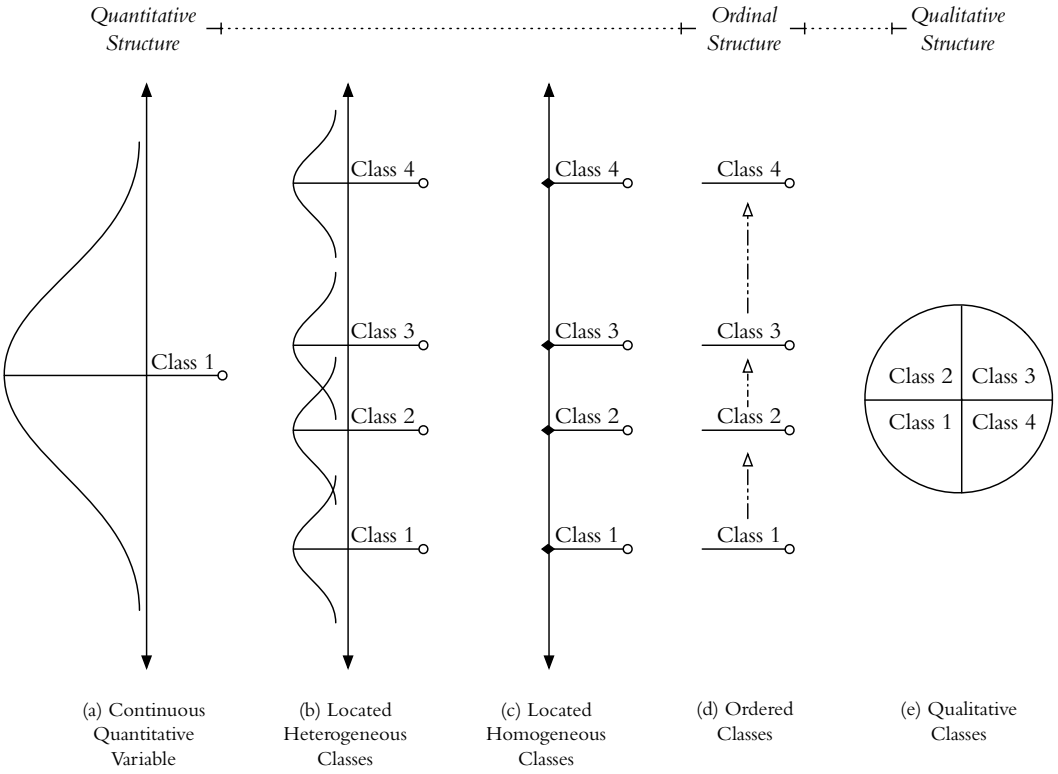


Figure 2.6: Alternative Models of a Latent Variable

The structure in pane (a) in Figure 2.6 is perhaps the most popular way of conceptualizing variables in psychology and, accordingly, one with an extensive line of research in the area of item response modeling (Birnbaum, 1968; Rasch, 1960/1980). However, scholars in latent variable modeling have proposed models for dealing with the other kinds of structures (usually under the label of mixture models) presented in panes (b) to (e): heterogeneous located latent classes (Lubke & Muthen, 2005; Muthen & Asparouhov, 2006), homogeneous located latent classes (Formann, 1995; Uebersax, 1993), ordinal latent classes (Croon, 1990, 1991, 2002; Van Onna, 2004), and of course, latent class analysis (Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968).

2.3.5 Metrology

As a discipline dedicated to measurement, metrology is not exclusively concerned with the theoretical foundations of measurement, but also with the practical aspects associated with maintaining the measurement infrastructure that supports 80% of global trade (Kaarls, 2007). Achieving coordination at this scale is not a trivial matter, and has prompted efforts of multiple international organizations concerned with measurement, including the International Bureau of Weights and Measures (BIPM) to establish the Joint Committee for Guides in Metrology[16]. This committee is responsible for the development and maintenance of two central reference documents: the *International Vocabulary of Metrology* (VIM for its acronym in French; Joint Committee for Guides in Metrology, 2012) and the *Guide to the Expression of Uncertainty in Measurement* (GUM; Joint Committee for Guides in Metrology, 2008).

The VIM states that its purpose is "to be a common reference for scientists and engineers—including physicists, chemists, medical scientists—as well as for both teachers and practitioners involved in planning or performing measurements, irrespective of the level of measurement uncertainty and irrespective of the field of application. It is also meant to be a reference for governmental and inter-governmental

[16]The member organizations are the International Bureau of Weights and Measures (BIPM), the International Electrotechnical Commission (IEC), the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), the International Organization for Standardization (ISO), the International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP), the International Organization of Legal Metrology (OIML), and the International Laboratory Accreditation Cooperation (ILAC).

bodies, trade associations, accreditation bodies, regulators, and professional societies" (p. 1). To achieve this, the VIM presents a consensus perspective, that while acknowledging differences between the areas of application, assumes that "there is no fundamental difference in the basic principles of measurement." Since it was first published in 1984, the VIM is now in its third edition in an attempt to remain a living document that responds to new areas of application and developments in metrology (Mari, 2014).

In its latest iteration, the VIM includes definitions of more than 140 concepts, which outline the basis for conducting measurements and communicating measurement results. A key concept that is defined is quantity, which is characterized as a "property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference," in opposition to nominal properties, which are "propert[ies] of a phenomenon, body, or substance, where the property has no magnitude." Up to this point, the VIM is consistent with the distinction between quantity and quality that was discussed in section 2.3.1; however, the VIM departs from this separation with the inclusion of ordinal properties as a type of quantity. An ordinal quantity is defined as a "quantity, defined by a conventional measurement procedure, for which a total ordering relation can be established, according to magnitude, with other quantities of the same kind, but for which no algebraic operations among those quantities exist."

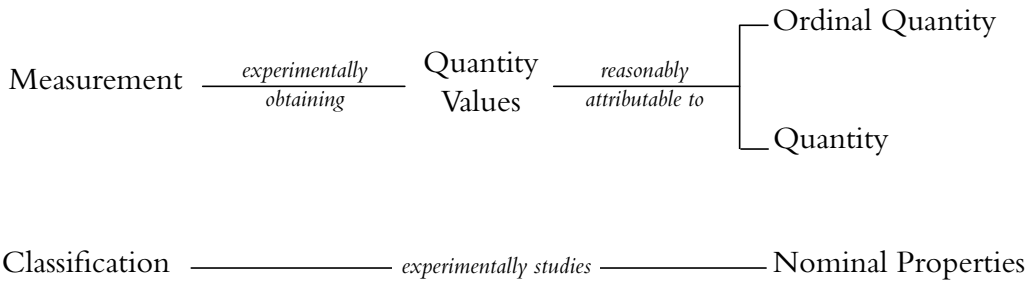


Figure 2.7: Measurement and Attributes in Metrology According to the VIM

With these three kinds of properties in the background (quantities, ordinal quantities and nominal properties), the VIM defines measurement as the "process of experimentally obtaining one or more quantity values that can reasonably be attributed

to a quantity" (p. 16), and accompanies this definition with three notes: (a) nominal properties cannot be measured (Figure 2.7), (b) measurement involves comparing quantities or counting entities, and (c) "measurement presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions."

The definition in the VIM is interesting when compared with the previous perspectives, as it emphasizes the experimental aspects involved in measurement, and while the definition may appear vague when referring to "reasonable attribution," the third note explicitly discusses conceptual and practical aspects that must be involved in measurement.

2.3.6 The Many Faces of Measurement

The previous sections presented a brief review of multiple definitions of measurement associated with several theoretical approaches, each emphasizing different aspects and answering to the varied demands of different domains of application. Any one of these definitions could be used to draw a distinction between who is *truly* doing measurement, and who is only pretending to do so, and depending on our preference, could divide the field between those who are on the right side of the definition and those who should be forced to renounce the use of the word "measurement."

However, regardless of these efforts to define or prescribe what measurement is and how it should be practiced, researchers and practitioners in physics, chemistry, engineering, medicine, epidemiology, psychiatry, psychology, sociology, political science, philosophy, and economics are routinely attempting to conduct measurement as part of their work, each one conducts their work under different theoretical traditions, but presumably, all aim to achieve some of the "precision and objectivity" (Porter, 1996, p. 23) associated with the concept of measurement. As Finkelstein (2003, p. 39) has remarked, "the requirements of those wider domains of knowledge and reasoning have led to the broadening of the concept of measurement, so as to encompass other forms of symbolic description."

In light of these varied ways in which measurement is practiced, Mari (2013) indicates that, "since in the scientific and technical literature multiple, sometimes

incompatible, definitions of 'measurement' can be found, identifying a single conceptual framework is a significant target for measurement science ..." (p. 2889). It is the search for this common conceptual framework that has prompted the development throughout history of different, and successively broader, characterizations of measurement in an effort to acknowledge new areas of application. The spirit of this search is nicely captured by Duncan (1984):

> And if theory of measurement, more broadly conceived than the theory of scale types, stops short of accounting for these measures which are among our main working tools, we must create a more comprehensive theory of social quantification. (p. 154)

It is worth noting that Duncan (1984) had serious reservations about Stevens' theory, and considered for instance that his inclusion of nominal scales as a level of measurement was a mistake. Nevertheless, as the above quote indicates, his commitment was to the creation and adoption of a definition of measurement that accounted for *our main working tools*.

However, not all scholars subscribe to this program, generating tensions between those who favor more encompassing definitions of measurement by considering socio-historical factors versus those who favor a time-invariant conception of measurement, cautioning about potential redefinitions as the work of "uncritical scientist[s] in a particular area [who] could, for socio-historical reasons, come to misunderstand a concept such as measurement and use it in ways inconsistent with its wider theoretical commitments" (Michell, 1999, p. 2).

The contrast between these two "poles" has naturally prompted some scholars to look for a middle ground, a third way so to speak, as can be seen in Savage (1970) where he contrasts *the narrow view*, which has a stringent definition rooted in classical notions that had served well the old physical sciences since Aristotle, and *the broad view* of measurement, which attempts to adapt and extend both the concept and practice of measurement. This push for flexibility in the latter has been associated with drive to serve the needs of the fledgling new psychological science and satisfy the demands that society at large increasingly demanded of it (cf. Hornstein, 1988; Michell, 1999), and of course, the lack of success found in the straight imitation of the measurement techniques of the physical sciences. Berka (1983) echoes this dilemma, asking:

> In what extension should we then conceive of the concept of measurement? Under the influence of the extensions of measurements in the social sciences, two different levels of measurement are usually differentiated: the so-called *lower* level (classification, numbering, numeral designation and ordering) and the so-called *higher* level, which corresponds to the scope of measurement as it is understood in the context of physical measurement. These different levels are also terminologically fixed by virtue of the differentiation between qualitative and quantitative measurement, or between classical and liberalized theories of measurement. In the same sense, we shall speak for the time being about the *wider* and *narrower* comprehensions of measurement. (p. 33)

Considering both of the extreme positions unacceptable, Savage expresses the desire to find a characterization of measurement that "is as broad as it can be without doing undue violence to the ordinary meaning or to the technical meaning of the term 'measurement' " (p. 197).

However, if at some point the movement towards a broader definition was spearheaded solely by researchers and practitioners in psychology and the social sciences, this is no longer the case. An interesting example of this lies in the point of contention between the CTM and the more liberal definitions and relates to the possibility of measuring ordinal attributes. As indicated in section 2.3.1, under the Aristotelian distinctions embraced by the classical conception of measurement, ordinal attributes are qualities, not quantities, and therefore cannot be measured, as measurement can only be performed when dealing with quantitative attributes (H. M. Johnson, 1936; Michell, 1997b, 1999, 2005). The representational measurement theory (Krantz et al., 1971/2007; Pfanzagl, 1968; Stevens, 1946) does not abide by the same restrictions, explicitly allowing the creation of ordinal measurement scales. Michell (2012b), perhaps the most prominent advocate in the last couple of decades for the classical, more stringent, perspective of measurement, considers that "merely ordinal attributes with impure differences of degree, [are] a feature logically incompatible with quantitative structure" (p. 255) and charges psychology and the social sciences with deluding themselves for attempting to measure such qualities. However, the acceptance of the possibility of ordinal measurement is in fact not restricted to psychology and the

social sciences, but is part of the conceptual toolbox of modern metrology, where the concept of *ordinal quantity* is officially recognized in the International Vocabulary of Metrology (Joint Committee for Guides in Metrology, 2012), effectively placing the International Bureau of Weights and Measures, home of the International System of Units, in disagreement with the classical definition of measurement.

The drive to separate classification, ordering and quantification in the current context also seems to ignore the distinction between the assumptions that we have about the attribute that is being modeled and the methods used to investigate it. This difference is important in light of the current use of statistics to model attributes of interest that are not only quantitative, but also ordered and qualitatively different. The confusion between the characterization of the methods and the hypothesized structure of the attribute of interest hinges on an important ambiguity in the meaning of "quantitative" and "quantity." We can see the conflation between the type of methods used to study an attribute of interest and the structure of said attribute in this quote by Michell (2012a):

> One of the defects of Stevens's well known classification of "scales of measurement" (Stevens, 1946) is that in assimilating classifications ("nominal scales" in Stevens's terms) and orderings ("ordinal scales") into his concept of measurement, the conceptual difference between the *qualitative methods of classifying and ordering* and the *quantitative method of measurement* is obscured [emphasis added]. The simplest way to see this difference is to note the fact that in "nominal" and "ordinal scaling" the use of numerals is optional because all of the information contained in such "scales" can be expressed non-numerically. For example, the classes comprising a classification can be given non-numerical names and the categories constituting an ordering can be designated using terms from any ordered series, such as letters of the alphabet. (p. 6)

In this quote, we can see how Michell is bundling the type of methodology used to study the attribute with the model of scale type or structure of the attribute, rejecting *a priori* the possibility of use *quantitative* methods for classification and ordering as well as *qualitative* methods for measuring. I contend that the current use of latent variable models to study latent classes and ordered latent classes does in fact allow us

to conduct the former, and as I will argue later in Section ??, the latter is an integral part of the process of studying "quantitative attributes."[17]

If we go back to what is considered one of the first definitions of a quantity, we will find that Euclid's definition is restricted to ratios of magnitudes (see Section 2.3.1). Later on, Hölder's axiomatization still considered quantity as ratios of magnitudes, but extended their application to intervals between points on a line (Hölder, 1901; Michell & Ernst, 1996, 1997). Similarly to Hölder, Stevens (1946) considered "quantitative" to apply to his interval and ratio scales. At the same time, latent variable modeling as a tradition is usually considered a "quantitative" methodology for its reliance on probability and statistics. On this regard, it is possible to see that simple transformations of probability (i.e. likelihood of an event expressed on a range from 0 to 1) will make them consistent with Hölder's axioms. Specifically, an odds transformation will make them conform to Hölder's axioms of quantity and a logit transformation will make them conform to Hölder's axioms for intervals (Freund, 2014).

In the use of a probabilistic statistical model, "quantitative" has a much broader scope than the definitions of quantitative by Euclid, Hölder or even Stevens (1946). Why? because within a framework such as LSA, those probabilistic statistical methods are not restricted to study attributes that are "quantitative" in Hölder's sense, and can be explicitly applied to study attributes that are hypothesized to be qualitative or ordinal and modeled through their effects observable variables. In other words, "quantitative methods" in that larger sense, when applied in a measurement context, do not presuppose that the latent variable is "quantitative" in the narrow sense.

In sum, the concept of measurement, and that of quantitative methods, is currently being used across the board in ways that do not necessarily conform to traditional (i.e. classical) definitions of measurement, pushing the boundaries of what constitutes "undue violence" to our technical understanding of it. Moreover, what constitutes a technical understanding of measurement, and the theoretical commitments that it entails will vary in different areas. In this context, disagreement on what is constitutive of measurement is bound to appear. For example, Mari (2013) in his

[17]It is important to clarify that I am in no way implying that we can fully examine qualities through quantitative methods nor quantities through qualitative methods. I am simply stating that the use of either methodology is in principle possible and can yield useful outcomes.

review of the literature, finds definitions that focus on the structure of the measurement process, the results of the measurement process, the comparison to a unit of reference, and the requirement of the correspondence to numbers.

Pragmatism offers here the advantage of being flexible and fallibilist, encouraging us to abandon the fantasy of a timeless and perfect definition that attempts to establish decontextualized/definitive demarcation criteria for what is truly measurement.

2.4 Measurement: Prototypes and Resemblances

It is my view that the attempt to establish a single demarcation criterion does not do justice to the range of current well accepted practice. On the one hand, it is unclear that we can "put the genie back in the bottle" and return to a more restricted understanding of measurement in order to rein in what is now a broad "ordinary meaning" in favor of a stringent and idealized "technical meaning." On the other hand, it seems legitimate to be concerned about misunderstanding and possible misrepresentation if we acknowledge that we have different communities using measurement in different ways. Is there a way to, acknowledge different practices and approaches as engaging in the overall endeavor of measurement, while at the same time acknowledging their differences? I think the answer is yes, and it involves recognizing that the search for a single definition of measurement is in a broader sense a problem of categorization, which is usually approached by attempting to provide a clear technical definition of the concept of measurement. Trying to characterize the meaning of a concept in this manner is consistent with what is called the *classical theory of concepts* as discussed in Section 2.2.1, and adopting Rosch's (1978, 1999) *prototype and graded structure account* might help in handling the multiplicity of definitions and approaches to measurement.

Laurence and Margolis (1999) present an example of how the classical approach characterizes a concept by using the example of the concept of *bachelor*:

> So BACHELOR might be composed of a set of representations such as IS NOT MARRIED, IS MALE, and IS AN ADULT (p. 9).

Based on such a definition, the classical theory predicts, (a) that categorization should be a clear cut matter (i.e. you either fulfill or do not fulfill the set of necessary

and sufficient conditions), (b) that all members of a category are equal members and (c) that we should be able to specify these conditions for concepts that we use (Medin, 1989).

However, empirical evidence contradicts these predictions, indicating that people encounter unclear and fuzzy cases (Laurence & Margolis, 1999; Medin, 1989). Similarly, it seems that within a given category some elements are indeed "more equal" than others. As Medin (1989) indicates when summarizing these findings, "people judge a robin to be a better example of bird than an ostrich is and can answer category membership questions more quickly for good examples than for poor examples" (p. 9).

A key feature of Rosch's account stems from its explicit recognition that for any given category, not every one of its members was considered to be equally representative of it; in her theory, she referred to the "best examples" of a category as *prototypes*. Many different factors can account for what makes a specific example prototypical of category, including elements varying from the frequency of the example to cultural and physiological factors (Rosch, 1999). For instance, what counts as a prototype can vary from one context to another, such as in the case of animals: "while dog or cat might be given as prototypical pet animals, lion or elephant are more likely to be given as prototypical circus animals" (Rosch, 1999, p. 67).

Finally, empirical results indicate that we are not as good at identifying the necessary and sufficient attributes required to define a category as the traditional view expected (Rosch & Mervis, 1975). By contrast "attributes appeared to have a family resemblance (Wittgenstein, 1953), rather than a necessary and sufficient structure." (Rosch, 1999, p. 67).

The notion of a relation of *family resemblance* between members of a category was famously illustrated by Wittgenstein (1953/2003) with the example of what counts as a "game":

> Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?—Don't say: "There must be something common, or they would not be called 'games'"—but look and see whether there is anything common to all. For if you look at them you will not

> see something that is common to all, but similarities, relationships, and a whole series of them at that. To repeat: don't think, but look!—Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball-games, much that is common is retained, but much is lost.—Are they all 'amusing'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! sometimes similarities of detail. And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear. And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities.
>
> I can think of no better expression to characterize these similarities than 'family resemblances'; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.—And I shall say: 'games' form a family. (§66-67; p. 27e-28e)

This notion is markedly different to a traditional view of categorization, which demands that all members of a category share the same set of necessary and sufficient attributes. Rosch follows Wittgenstein in arguing that our concepts are extended "as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres" (Wittgenstein, 1953/2003, p. 28e).

2.4.1 Prototypical Measurement Practices

I think that the framework provided by the prototype/graded structure research program, and Wittgenstein's notion of family resemblance, can help us organize and clarify the multiple relations and differences between alternative approaches to characterizing measurement.

From this perspective, it is possible to recognize the role of more traditional approaches to measurement as prototypical instances of measurement (e.g. determining the length of a table with a ruler), without restricting the concept to them. As mentioned above, this framework also recognizes that which prototype will be considered as representative of concept depends on the context, which can also help reconcile that "length and a ruler" can come to mind when talking casually about measurement, and "math knowledge and a quiz" when talking about teachers and classrooms or "literacy and PISA" when discussing educational policy.

To show how thinking about measurement definitions in this way can help, consider the diagram in Figure 2.8. This diagram is meant to represent different measurement practices according to different communities within the larger, amoeba-shaped, scope of the practices that use the concept of, and purport to conduct, "measurement." Within this area, the diagram highlights seven points, each representing a prototypical measurement practice according to the theories reviewed earlier.

We can begin reviewing this diagram with one prototypical measurement practice, the classical theory of measurement, labeled M_{CTM} and located more or less at the left center, which historically constituted a most influential prototype of the notion of measurement. This prototype is surrounded by other measurement practices that have emerged as the use of the concept has changed and adapted to the needs and use cases of different communities of researchers or users, including in this case metrology M_{Met}, the representational measurement theory M_{RMT}, axiomatic measurement theory M_{AMT}, operationalism M_{Oper}, Stevens' typology of scale types M_{TST}, and latent variable modeling M_{LVM}.

As the meaning of measurement is expanded, some features are dropped and others preserved, maintaining perhaps only a family resemblance relation to each other. Faced with this situation, there are those who advocate a specific practice and say "this is the true definition of measurement," and thus can correctly point out that the other practices do not have the set of attributes that they deem necessary. Based

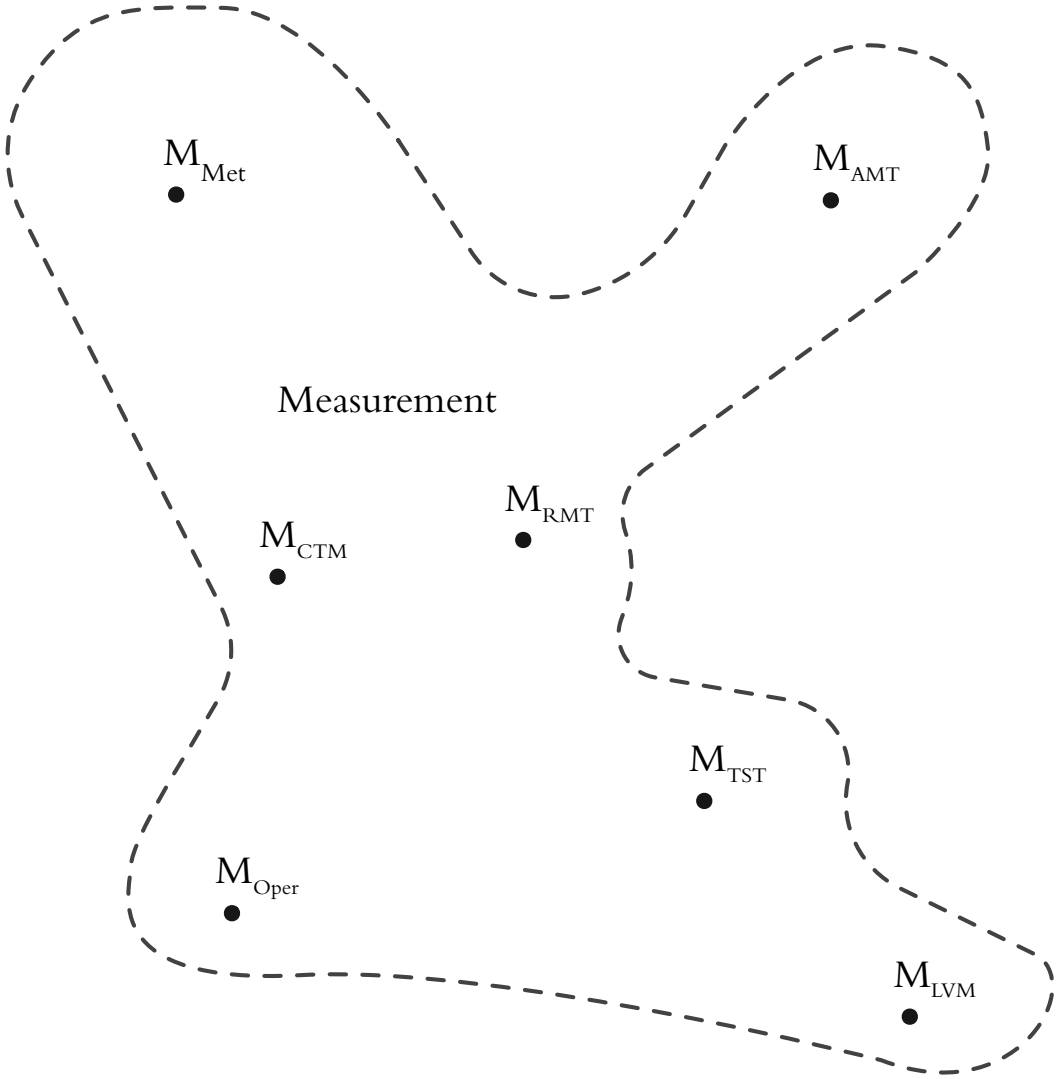


Figure 2.8: Multiple Prototypes of Measurement.

on this issue, those who favor the practice of M_{CTM} as the prototype of measurement want everyone else to stop using the concept of measurement and refer to what they are doing as something else. This is a most stringent way that we could approach the definition of measurement and it seems to me an option that is unhelpful in promoting good practices across the whole amoeba-like area.

On the other extreme, there are those who claim that measurement is the entire amoeba-shaped area, and under that premise, claim that all practices are equal members, all benefiting from the credibility that emanates from the concept of measurement. On this end, instances of measurement under Dingle (1950)'s radical operationalism would have to be considered as equivalent to any other measurement conducted under axiomatic measurement theory, latent variable modeling, or metrological procedure. In other words, and under the idea that measurement is an all-or-nothing proposition, everything we allow ourselves to consider them as "measurement", becomes and equal member in the classification: a 22 year old unmarried college student and the Pope are both bachelors after all. This is the most liberal way to approach to understanding measurement and it seems to me equally unreasonable.

A different issue may arise in between these extremes if some measurement practices under one prototypes (e.g. M_{TST}), are being mistakenly identified by the public as practicing measurement according to a different prototype, (M_{Met}), either by miscommunication, or worse, by misrepresentation.

The account that I propose emphasizes the fact that these alternate measurement practices are indeed different in more ways than one, and while they all embrace the concept of measurement, it is unwarranted to simply assume that their different measurement practices will underwrite claims with the same precision. By acknowledging a more complex structure to the use of the concept of measurement we can abandon the all-or-nothing propositions that, on one extreme would prohibit uttering the word "measurement" outside the classical understanding[18], and on the other would attempt to establish all measurement practices as equally precise and established.

The large amount of white inside the space amoeba-like area in the diagram also hints at another aspect of measurement practice that is worth highlighting, namely,

[18] Or metrological if its practitioners suddenly start advocating for their right to the throne of the kingdom of measurement.

that although we can identify a certain set of ideal prototypes under each approach (the black points in this case), most of the work on measurement occurs in the space between of these ideal types (not prototypical), but still within the realm of measurement. In other words, when scholars engage in discussions about measurement or when practitioners conduct measurements, they do so under the influence of multiple approaches. I would go as far as arguing that what is considered "measurement" for those who use the concept but do not study it directly is likely nothing more than the collection of practices of the prototypes that are closer to their area of application, which are usually adopted without explicit mention or allegiance to one or other of these prototypes.

2.5 A Pragmatic Definition?

I have argued for the idea that the multiple measurement practices discussed earlier are related by a set of family resemblances to each other, where each measurement prototype highlights different aspects depending on its emphasis and context of application. Framing the myriad of definitions and theories as different prototypes, with shared family resemblances to each other, allows us to consider them in a complementary way instead of thinking of them as fully opposed, competing after the position of single, true definition.

I have also argued that because of this set of resemblances, attempting to define measurement with a single and clear cut definition might not be the best strategy if we are committed to developing a comprehensive understanding, and improvement, of measurement across different disciplines. This is not meant to dismiss the attempt to characterize measurement through definitions and theories, but to demand less of them. In particular, to stop expecting them to present an exhaustive characterization with clear cut limits, and think of them more as maps that allow us to highlight relevant elements of the practice of measurement under the varied demands that emerge in different areas of application.

In this context, I would like to present a Pragmatic Perspective of Measurement (PPM) to highlight what I see as some key commonalities that are shared by most of the measurement approaches previously discussed. By leveraging the emphasis of Pragmatism on the ideas presented in Section 2.2, I aim to present a characterization

of measurement that highlights some of the family resemblances, without pretending to be exhaustive or prescriptive.

Pragmatism offers an interesting alternative to a core assumption shared by some of the major accounts of measurement, namely the classical and representational theories of measurement. Despite their differences, these theories assume that there is a true structure, either of the attribute or the qualitative empirical relations, that we are trying to recover or replicate. Accordingly, a measurement is good to the extent that it resembles this true structure and approximates true values. Pragmatism invites us to look for an alternative conception of measurement that eschews this copying metaphor, and with it, a set of theoretical and philosophical problems related to the definition and demarcation of measurement.

When thinking on how to characterize a domain of knowledge from a Pragmatic perspective, I find James' (1907/1995) example of a radical Pragmatic definition of physics by W.S. Franklin (1903) illuminating:

> I think that the sickliest notion of physics, even if a student gets it, is that it is "the science of masses, molecules and the ether." And I think that the healthiest notion, even if a student does not wholly get it, is that physics is *the science of the ways of taking hold of bodies and pushing them*! [emphasis added] (p. 15).

In order to develop a Pragmatic account of measurement, I would propose the following characterization of measurement in the same spirit as Franklin's definition of physics (1903):

> Measurement is (i) an activity of classification, ordination, or quantification of a set of elements (ii) according to a model (iii) of a relevant attribute (iv) in service of a larger goal.

2.5.1 Why an activity? Why three of them?

Noting that measurement is an activity is not unique to this definition, as other definitions implicitly characterize measurement as an activity, either as *numerical estimation* (Michell, 1997b, p.383), *assignment of numerals* both by Campbell (1920, p. 267)

and Stevens (1946, p. 677), or *obtaining and reasonably attributing* in the case of the VIM. The difference here lies simply in making it explicit to remind us that measurement is being conducted to engage with the world, not to passively mirror the world under the spectator theory of knowledge, but to support a successful use. Viewed in this light, a measurement need not be understood as replicating or recovering a set of *true* values, but as actively classifying, ordering and quantifying.

In this proposed definition, I have included not one, but three types of activities: classification, ordering and quantification. This is of course not a trivial issue, as a large part of the debate over the definition of measurement is related to this point.

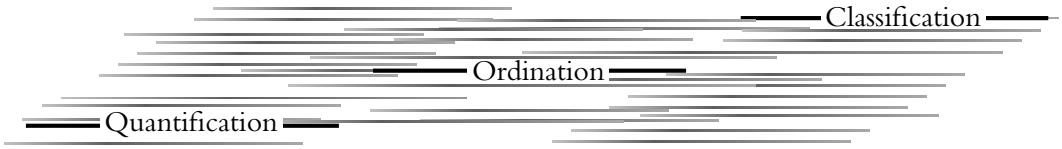


Figure 2.9: Three Fibres in the Thread of Measurement

To explain why it might be relevant to list these specific activities, it is important to consider the *level of the activities*. A simple example will suffice to clarify it. Let us take as an example "having dinner." There are many actions that could be used to describe what is occurring. While some may say I am having dinner, for instance, we could say more broadly that I am eating, that I am acquiring nutrients, or even that I am subsisting. We could, on the other hand, say more narrowly that I am sitting down, that I am reaching for food with my hands, or that I am chewing and swallowing. In this example, I think that "subsisting" is, for most purposes, too broad a level of description, and that "chewing and swallowing" is too narrow. I would argue that one of the main sources of confusion when discussing measurement is the different levels at which the activity of measurement is described, with most definitions focusing on too concrete a level by talking about "estimation," "assignment," "attribution," versus much more abstract alternative characterization such as "a central epistemic enterprise" (Trout, 1998, p. 150).

I think that a more descriptive way of characterizing what we are attempting to achieve when we purport to conduct measurement is to describe measurement in between these two levels, hence classification, ordering and quantification. The verbs more directly describe the outcomes that researchers and practitioners have in

mind when they attempt to conduct a measurement. Distinguishing between these activities helps clarify whether we want to distinguishing differences of kind, order or quantity in the relevant attribute that we are measuring.

There are of course potentially infinite levels of distinctions that could be made regarding the activities that we engage in when measuring. However, it seems to me that the debate over the practice of measurement is concentrated on two main levels that are usually brought to bear in discussion regarding the definition of measurement. These levels are presented in Figure 2.10.

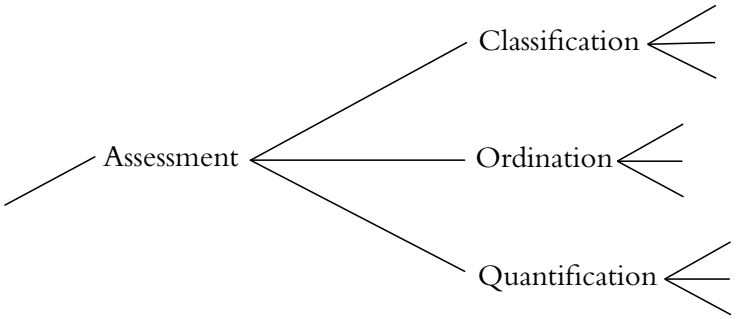


Figure 2.10: Two Possible Uses of "Measurement"

On the left side Figure 2.10, we have a higher level activity, assessment, while on the right hand side, we have another level with three, more concrete, activities: classification, ordering and quantification. The lines in this diagram are meant to indicate not only that assessment is a higher level activity than the other three, but also that assessment can be considered as a lower level concept of another activity (e.g. epistemic enterprise), and that each one of the lower level activities included in the diagram could in their own right be analyzed in lower level actions (e.g. estimation, assignment).[19]

Measurement has purposefully been left out of this diagram because an important part of the debates regarding the concept of measurement is due to the conflict between approaches that attempt to establish measurement at the level comparable to assessment (e.g. most representational approaches), versus those that want to restrict

[19] What about counting? From this PPM, there would not be an Aristotelian distinction between discrete and continuous quantities, therefore there would be no need to exclude a priori counting as a form of measurement

to quantification (CTM), and perhaps, ordering (e.g. the current definition under the VIM). When viewed in this light, it is possible to think that we could avoid much disagreement by simply abandoning the word "measurement," and settling for the names in Figure 2.10 or other similar ones; however, as there is considerable prestige in the name "measurement," it seems unlikely that researchers all over the world will simply decide to rechristen their practices.

However, as can be inferred from the inclusion of classification, ordering and quantification in the definition above, if suddenly the entire world where to decide in this renaming effort, the title of this paper could be simply be renamed to "A Pragmatic Perspective of Assessment."[20]

Why am I inclined towards a view of measurement as closer to assessment, or at the very least as broader than quantification? Because the attributes that interest us in the social sciences can be, in principle, modeled in any of these three ways, and I think that our study in all of these cases would benefit from a common framework that aspires to produce reliable and actionable knowledge. Purposefully prescribing walls between these traditions, and attempting to enforce them, is counterproductive to this agenda.

When studying an attribute of interest in the social sciences, such as proficiency in mathematics, we traditionally hypothesize that inter-individual differences can be well modeled in terms of quantitative differences. However, it is also possible to model these differences in terms of ordered levels of proficiency, or even, in some cases, as different kinds of approaches (e.g. solution strategies) adopted by the respondents. Each one of these models has different implications regarding the structure of the attribute of interest and different possible inferences that we can make about the persons that are being studied. When working to understand an attribute, we should avoid forcing *a priori* the selection of a continuous, ordinal or categorical model.

The alternative is to emphasize the divide between them, developing a discipline of measurement, with its vocabulary and institutional knowledge solely for the scenario in which we want to model the attribute as quantitative. It is to have a parallel

[20]One difficulty not often discussed among proponents of the rechristening of some practices as "assessment" instead of "measurement" is the assumption that a comparable word at that level of generality exists in every language. This is by no means a given, and at least in Spanish, there is no separate word for "assessment" aside from words like *medición*, *evaluación*, and *valoración* that respectively would correspond to *measurement*, *evaluation*, and *valuation*.

methodological tradition solely dedicated to ordinal modeling and yet another one for classification, each one with its own lingo. I would argue that this has been the alternative favored so far. We have a psychological and educational measurement tradition dominated by the focus on modeling attributes as quantitative (Birnbaum, 1968; Rasch, 1960/1980), a separate tradition of practice that focuses on ordinal measures (Cliff & Keats, 2003), and yet another methodological tradition and professional community focused on classification (Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968). One possible, perhaps ideal, outcome of having these separate traditions could be that researchers in the social sciences could take advantage of the methodological expertise of any and all of them, freely exploring the alternative ways in which the attribute that they are studying could be structured. However, I do not think this is occurring. On the contrary, it seems that the association of the concept of measurement in the social sciences with one of these practices, quantification, in addition to the separation of these three traditions, hampers the flexibility of social scientists to consider alternative modeling practices of the attributes of interest. If one is taught classification methods, everything becomes a taxonomy; if one is taught continuous measurement models, everything looks like a number line.

To be clear, I do not think that separation between these research traditions is *per se* negative, or that specialization is a negative development. However, I do think that it is possible to do a better job relating these traditions, highlighting their family resemblances, and not limiting the way in which we conceptualize the attributes we study due to barriers in our methodological traditions. This is especially necessary as our statistical frameworks blur the distinction between these traditions. In this context, I think that understanding the concept of measurement at the level of "assessment" in Figure 2.10, is a step in the direction of more methodological and conceptual flexibility in the social sciences.

2.5.2 Why according to a model?

Another important aspect of the proposed definition makes explicit that the classification, ordering or quantification of the elements of a set of elements is conducted *according a model* of the attribute of interest. This model, usually a substantive scientific theory, is key, as it defines which objects can be conceived as having the attribute,

establishes what kind of relations characterize the attribute, guides our expectations regarding potential study and manipulation (e.g. does it make sense to order it? can we calculate distances between different instances of the attribute?), and finally, provides the basis for justifying why the instrument or transducer[21] is an appropriate indicator of the attribute of interest.

One of the critical aspects that is played by the model that we have of our attribute of interest is that it will constraint the level of precision to which we can interpret the results of our measurements. As Boring (1920) emphasized in the early 20th century:

> But, if in psychology we must deal—and it seems we must—with abilities, capacities, dispositions and tendencies, the nature of which we can not accurately define, then *it is senseless to seek in the logical process of mathematical elaboration a psychologically significant precision that was not present in the psychological setting of the problem* [emphasis added]. Just as ignorance will not breed knowledge, so inaccuracy of definition will never yield precision of result. (p. 33)

A warning worth keeping in mind in the social sciences where concepts often seem to be—echoing Allport (1935) in his seminal discussion of attitudes—"measured more successfully than they are defined" (p. 828). A situation that should prompt researchers and practitioners to wonder about the criterion used for defining success.

More generally, the "theory-ladenness" of measurement, meaning "the pervasive use of theoretical assumptions in designing measurement apparatuses and interpreting their indications" (Tal, 2013, pp. 1167–1168), and its complexities have been extensively discussed (see for instance Chang, 1995; Kuhn, 1961; Van Fraassen, 2008), such that "contemporary scholarship has come to view measurement as theory dependent by default" (Tal, 2013, p. 1168).

As Tal (2013) has pointed out, "contemporary studies show that theoretical and statistical models of a measurement process are necessary preconditions for obtaining meaningful measurement outcomes in the first place" (p. 1166). The role of models as part of measurement is featured across different measurement traditions, including

[21] A measuring transducer is defined in the VIM as a devise, used in measurement, that provides an output quantity having a specified relation to the input quantity (Joint Committee for Guides in Metrology, 2012).

the latent variable literature in general (Skrondal & Rabe-Hesketh, 2004), and psychometrics in particular (Wilson, 2013), as well as economics (Boumans, 2001) and metrology (Mari, 2003).

However, the inclusion of models of the attribute and models of the measurement process could raise a potential contradiction with the original Pragmatic emphasis on the abandonment of the mirroring metaphor for the practice of measurement. This contradiction would stem from maintaining the attachment to the spectator theory of knowledge and its commitment to the idea that models are supposed to be evaluated in terms of how closely they resemble reality. Teller (2001) describes this view of models, stating that "the photograph provides a good icon: The ambition has been to produce a perfect likeness of nature, a perfect model." So, to recap, under this traditional spectator theory of knowledge we have a natural kind that is being represented (i.e. mirrored) by a model in one or more ways, where the adequacy of the model is judged in terms of how close our model resembles the aforementioned natural kind. However, we need not conceive models this way.

The next section will tackle an alternative to the notion of natural kind in favor of relevant ones, so the present question is, how can we judge the adequacy of a model in a way that does not depend on the idea of resemblance.

An an example, consider the case in which you are interested in judging the quality of a set of models for the human heart. You are presented with four different models:

- The first one was created as an aid to explain hearts to middle school students.
- The second was created as an aid in training surgeons.
- The third model was created as a specification for creating an artificial heart.
- And the last model was created as a specification for cloning a heart replacement.

Does it make sense to ask which of these representation is the correct one? I think not. Does it make sense to answer this based on similarity with a canonical heart (whatever that means)? They all have some similarities, but none of them mirrors all of them.

The idea here is that even when modeling ostensibly the same object, a human heart, the idea that we can resort to resemblance as a criterion for model adequacy is

a misleading one, as what counts as features that are to be mirrored in the model will vary significantly depending on the use that we want to give to the model. Additionally, it is unclear that a perfect reproduction would constitute a useful alternative.

Consider as another example the case of the attribute *college readiness*. We might consider this attribute, and be concerned about modeling it, because we do not want to send to college (nor fund) students that we think might not succeed there or finish their education. Alternatively, we might be concerned with the limited number of vacancies, and we want to make sure that those who can take the most advantage of them should be the ones selected to attend. Based on these or other concerns, we could think of multiple models of how this attribute may vary across individuals.

If our goal is selecting only those who are deemed "ready," regardless of the fact that those may be less than the amount of vacancies, we may think of college readiness as a binary attribute, such that any given person is either ready or not ready to succeed in college. Under this model, we can conceive of a qualitatively distinct class, the college-ready class of students, that is characterized by distinct academic, emotional and social attributes. We would therefore we want to classify students as being members or not of this class.

If our goal is not only to select those who are ready, but also to offer different kinds of support activities targeted to students' strengths and weaknesses, we may think that college readiness is best modeled by multiple classes. Under this model, we can hypothesize that there is a first class of students who are ready for college because of their academic prowess, a second class of students whose college readiness is based on their emotional maturity, a third class that is characterized by good social skills, and finally, a fourth class of students that are not characterized by any of the previous profiles.

If all we want to do is fill the available vacancies, without any assumptions regarding the "absolute" level of readiness of the students, we could use an ordinal model. In this case we think that some students are overall "more ready" or "less ready" than others, and it is possible to order or rank them from 1 to N based on these more-or-less relations, and then count until we run out of vacancies.

Finally, if we think that there is a specific cutpoint that we consider to be "ready enough" (just like the height check before a roller coaster), we may think of college readiness as a continuous variable, very much like height, where we see each

students' readiness as summarized in a single metric that allows quantitative comparisons among them. Alternatively, this last model would be very useful if, for instance, we had a well established function for manipulating college readiness, such that we could estimate the amount of additional hours/resources of training that would be required for the students to achieve such cutpoint.

Each one of these descriptions of the relevant attribute "college readiness" are in principle possible, and it is an empirical question whether they prove useful or not when put to work in a given context. It might be the case that more than one of these models proves fruitful, or perhaps none of them will. Nevertheless, what we are doing in these different cases is trying to classify, order or quantify inter-individual differences according to our model of this relevant attribute to accomplish a larger goal, as opposed to focusing on whether college readiness is truly a natural attribute, or if we should ban or not some of these models because they do not adequately capture the true nature of college readiness.

This point is more generally made by Rorty (1999) about objects, but is equally valid for attributes[22]:

> No description of an object is more a description of the 'real', as opposed to the 'apparent', object than any other, nor are any of them descriptions of, so to speak, the object's relation to itself—of its identity with its own essence. Some of them are, to be sure, better descriptions than others. But this betterness is a matter of being more useful tools—tools which accomplish some human purpose better than do competing descriptions. All these purposes are, from a philosophical as opposed to a practical point of view, on a par. There is no over-riding purpose called 'discovering the truth' which takes precedence. As I have said before, pragmatists do not think that truth is the aim of inquiry. The aim of inquiry is utility, and there are as many different useful tools as there are purposes to be served. (p. 54)

[22]Especially given Rorty's panrelational stance, where absent the idea of essential features, "there is nothing to be known about [objects] except an initially large, and forever expandable, web of relations to other objects. Everything that can serve as the term of a relation can be dissolved into another set of relations, and so on for ever. There are, so to speak, relations all the way down, all the way up, and all the way out in every direction: you never reach something which is not just one more nexus of relations." (p. 53-54)

In this sense, a model, like a theory and language in general, is also "an adaptation to the environment, a set of tools for dealing with the causal pressures exerted by the environment, rather than a way of mirroring it" (Tartaglia, 2007, p. 214). As such, determining the adequacy of a model of a heart, a model of college readiness, or any other object or process, will depend on our purpose given a context of application. This point has also been raised by Minsky (1968), where he defines model by stating:

> To an observer B, an object A^* is a model of an object A to the extent that B can use A^* to answer questions that interest him about A. The model relation is inherently ternary. *Any attempt to suppress the role of the intentions of the investigator B leads to circular definitions or to ambiguities about "essential features" and the like.* [Emphasis added] (p. 426)

Echoing the words of Dewey, the quality of our model can be evaluated based on how good the model is at doing the kind of work that we need it to do for us.

This perspective offers us an amended version of Box' (1987) dictum "all models are wrong, but some are useful" (p. 424), which we could revise to indicate simply that *some models are useful*.[23]

2.5.3 Why a "relevant" attribute?

Another aspect of the proposed definition that is worth highlighting is the explicit reference to *relevant attributes*. According to Webster's Third New International Dictionary, "relevant" is defined as:

> Bearing upon or properly applying to the matter at hand; affording evidence tending to prove or disprove the matters at issue or under discussion; PERTINENT

Relevant is used in this sense here to make a distinction from the idea of a *natural* attribute, either a kind and/or an universal. This distinction is in the same spirit as Nelson Goodman's (1975) distinction between *natural* and *relevant* kinds:

[23] Another revision to Box' famous quote, made in an akin spirit but from a different line of argumentation, was posed by Tarpey (2009), who proposes that "All models are right, [but] most are useless"

> I say 'relevant' rather than 'natural' for two reasons: first, 'natural' is an inapt term to cover not only biological species but such artificial kinds as musical works, psychological experiments, and types of machinery; and second, 'natural' suggests some absolute categorical or psychological priority, while the kinds in question are rather habitual or traditional or devised for a new purpose. (p. 63)

This distinction is apropos here to present an alternative to the idea that measurement must be concerned with natural kinds or universals. Metaphysical realism, where the commitment is to natural kinds and/or universals as opposed to relevant kinds and relevant attributes, is illustrated by Michell (2011, p. 252):

> Concepts are not tools, somehow already present in our minds and able to be imposed upon a non-conceptual reality, for good or ill. Rather, concepts are features of reality. For example, when I see that this rose is red, whatever it is *to be a rose* and *to be red* (i.e., the concepts involved) are real qualities, present in indefinitely many particular plants and indefinitely many physical objects. Otherwise I could never judge veridically that this rose is red—something I can surely do.

As it was discussed in Section 2.3.1 Michell's take on how attributes are part of reality is a central feature of the CTM, and his example about color illustrates the extent to which he thinks of science in general and measurement in particular as discovering reality itself, describing it in its natural state, untainted by human activity. For this reason, it is interesting to make a small aside and focus on an alternative to this "let's just carve nature at its joints" approach to color in particular and attributes in general.

Despite Michell's self-confidence in his capacity to correctly identify the colors as they are in reality itself, not everyone agrees with this bullish assessment[24]:

> Color concepts are "interactional"; they arise from the interactions of our bodies, our brains, the reflective properties of objects, and electromagnetic radiation. Colors are not objective; there is in the grass or the

[24] On a personal note, being "colorblind" myself I have learned the hard way not to make those kinds of statements about color.

> sky no greenness or blueness independent of retinas, color cones, neural circuitry, and brains. Nor are colors purely subjective; they are neither a figment of our imaginations nor spontaneous creations of our brains. (Lakoff & Johnson, 1999, p. 24-25)

Varela et al. (1991) examine in detail the issue of color as a cognitive phenomenon, discussing some of its neurological, physical, and conceptual aspects. In their review, they dispute the common assumption that there *must* be some physical correlate of colors, pointing out that:

> This relative independence of perceived color from locally reflected light has been known to vision scientists for quite some time. The independence is manifested in two complementary phenomena. In the first, the perceived colors of things remain relatively constant despite large changes in the illumination. This phenomenon is known as *approximate color constancy*. In the second, two areas that reflect the same spectral composition can be seen to have different colors depending on the surrounding in which they are placed. This phenomenon is known as *simultaneous color contrast* or *chromatic induction*. (p. 160)

Varela et al. go on to conclude their examination of color by indicating that "we can now appreciate, then, how color provides a paradigm of a cognitive domain that is neither pregiven nor represented but rather experiential and enacted" (p. 171). As Rorty (1999) points out:

> To say that X is really blue even though it appears yellow from a certain angle and under a certain light, is to say that the sentence 'X is blue' is more useful—that is, can be employed more frequently—than the sentence 'X is yellow.' The latter sentence is useful only for occasional, evanescent purposes. (p. 51)

This aside about color is simply to illustrate a central aspect of the proposed PPM, and how it differs not only from the CTM but from all other approaches that, as characterized by Varela et al., emphasize the idea that in general "the world can be divided into regions of discrete elements and tasks", and, based on that premise, assert

that "[cognition] must, if it is to be successful, respect the elements, properties and relations within these pregiven regions" (Varela et al., 1991, p. 147).

From a PPM, the emphasis is on the idea of relevant attributes, and relevant objects for that matter, that emerge from our practices and efforts to cope with the world. If you are wondering at this point about who determines what is relevant, Hacking (2007) answers, directly in relation to Goodman's concept of relevant kinds, "to us, or to you, or to them, to those who group items together, for this or that purpose. Relevance is all there is to be said, in general" (p. 46).

2.5.4 Why in service of a larger goal?

The final, and perhaps crucial point raised in this proposed definition is the explicit characterization of measurement as a purposeful activity in service of a larger goal, highlighting that when engaged in measurement, we aim to accomplish something.

The reason for explicitly including goals in the definition is that, from a Pragmatic point of view, the goals provide the criteria for judging whether the measurement we are examining is a good one or a bad one depending on the extent that they help us accomplish them. Associating a measurement with a specific and explicit goal both (a) restricts its scope, hopefully tempering inferences and statements based on it, and (b) establishes the criteria by which the measurement is to be judged.

From a PPM, conducting a measurement is not an all-or-nothing proposition. Accepting the use of a procedure as a "measurement" is not an act of canonization, that ensures that its results are always true, that it can be used in any context or that we can draw any inference from it. The question is not *is this truly measurement?*, but *is this a useful measurement for this explicit purpose?*

Although the inclusion of goals as an explicit part of the proposed pragmatic definition of measurement can seem controversial, the role of goals in measurement has traditionally been part of the discussion, albeit on the background.[25] Take for instance Campbell's position on the matter:

[25] An example of this controversy was the debate generated by the introduction of *Consequential Validity* as part of the 1999 Standards for Educational and Psychological Testing. See section 2.6.1 for a discussion on this point.

> Measurement is only a means to an end; we want to express the properties of systems by numerals only because we are thereby enabled to state laws about them. (Campbell, 1920, p. 328)

And, more recently, Michell (2008c, p. 137), discussing Denny Borsboom's book *Measuring the Mind*:

> This does not distinguish measurement from cognition generally, and it misses the main point of measurement, which is to get to know about *quantitative* attributes.

I contend that, while Campbell's and Michell's proposed definitions of measurement do not explicitly reference the goals of measurement as an activity, these quotes indicate that the goals that they consider for the practice of measurement are indeed strongly shaping their perspective on the matter. Campbell's stated goals conceive measurement as a means to formulate laws of nature, and Michell presents measurement as solely concerned with learning about quantities, in accordance with his alignment with the spectator theory of knowledge. Campbell places quite an onus on those that want to engage in measurement, you either come up with natural laws or go home. For Michell and his realist approach to measurement, the main aim is to address the scientific task of measurement, which centers on determining if something really is or not a quantity, a daunting and very specific proposition. The commonality between the two is that measurement is conceived as a purely research oriented enterprise and a very narrow one at that. Their characterizations of measurement are in service of restricting measurement exclusively to those practices that help them address those goals. Our definition of the goals of measurement will certainly constrain our definition of measurement, and more importantly, it will prime our interpretation of what we *should* expect from its outcomes. In the case of these narrow goals, the constraint seems to be: *measurement if and only if we make inferences about nature's laws*.

These are of course valid goals, but they are only *some* of the possible goals, not in principle better or worse than other goals. I contend that this appraisal of the goals of measurement is simply too narrow, excluding valuable reasons for attempting to measure, and I am not alone on this. In his paper, *On The Nature and Purpose of*

Measurement, Ernst E. W. Adams (1966) presents a Pragmatic friendly—though not explicitly so—take on this matter, contesting the focus on research as the sole purpose of measurement:

> My position emphasizes the role of measurement in *application*, where a theory may be involved, but the objective is to do something or anticipate something, and not to create or test a theory. A comprehensive theory of science should give both aspects of it their due, of course, yet it seems to me that much recent writing greatly overemphasizes the "inquiry" or research aspect perhaps derogating the applied aspect as "mere engineering". (p. 153)

The drift from conceiving measurement as a practical problem solving tool towards one that focused on theory building has also been noted by Duncan (1984, p. 2):

> What I am trying to suggest is that many—and perhaps the most basic—of the procedures natural and social scientists use in measuring were actually invented to solve practical problems. In the beginning, measurement served social purposes only. The scientist may come into the picture when there is a recognized need to improve the measuring instrument. Or, taking the current practice of measurement as his point of departure, he may let his imagination work freely on ideas of amount, extent, magnitude, intensity, duration, numerousness, dimension, scale, and proportion to create abstract conceptual structures and systems of relationships.

The connection with the problem solving focus of measurement is key in the context of this Pragmatic framework because our goals are, in the most general sense, an expression of our need to adapt to the "causal pressures exerted by the environment" (Tartaglia, 2007, p. 214).

This is, of course, not the first appeal to usefulness in relation to a goal as a criterion for measurement. For instance Churchman (1959)—in a piece called *Why measure?*—presents a similar perspective: "Suppose, then, we propose that the function of measurement is to develop a method for generating a class of information

that will be useful in a wide variety of problems and situations" (p. 84). From this proposal, he indicates:

> We can begin by noting one rather striking consequence of the proposal. The objective of measurement can be accomplished in a number of ways... The qualitative consignment of objects to classes and the assignment of numbers to objects are two means at the disposal of the measurer for generating broadly applicable information. But which means is better? The striking consequence of the proposal is that measurement is a decision making activity, and, as such, is to be evaluated by decision making criteria. (p. 84)

A few years later, E. W. Adams (1966) would similarly argue for the importance of the purpose of measurement (while making a case against the abstract focus of the representational theory of measurement):

> Though the work of Campbell and others in the representational tradition has contributed to our understanding of measurement, the proponents of this approach have neglected to consider what it is that measurements are made *for*, and in so doing have been led to conclusions as to what measurement *ought* to be which are in serious disagreement with what scientists do. (E. W. Adams, 1966, p. 125)

E. W. Adams (1966) put forward what he called an "informational account of measurement" such that:

> ...measures of a quantity are not so much "true" or "false" as they are more or less informative about the phenomena which they are supposed to be indices of.

I, of course, think that E. W. Adams (1966) is onto something here, and I would simply suggest amending this to indicate that measurements are *more or less useful to the goal that we are trying to accomplish*. I think this modification would certainly be in the spirit of Adams' own treatment of measurement, as his discussion of the the Mohs scale of hardness illustrates: "The two basic things to know about this measure are:

(1) what it is used for, and (2) how it is made" (E. W. Adams, 1966, p. 133). These are two questions that, from this Pragmatic Perspective of Measurement, everyone that purports to measure should answer about their measurement instruments.

Having argued for the importance of considering the larger goals of measurement, it is good to tackle directly some common objections to the inclusion of a Pragmatic criterion such as this. A first concern is that considering a Pragmatic criterion based on usefulness in relation to goals will endorse irrelevant goals. Take for instance Michell's objection on this regard:

> Utility is a relative notion, and not all kinds of utility are relevant to the scientific enterprise. For example, there are a number of senses in which quantitative methods can be said to work well in psychology. First, those who have employed them have often profited from doing so. The public is impressed, careers are advanced, and money is made. Second, some quantitative procedures have been found to enable moderately useful predictions of other events. However, criteria such as these are really irrelevant to the scientific enterprise. The only criterion of interest to the scientific enterprise as such is that of advancing investigations into the ways of working of things. (Michell, 2003, p. 47)

Two points come to mind. First, the inclusion of goals as proposed by this PPM would explicitly deal with the case where someone purports to measure and claims that it is useful because he or she was able to get peer reviewed publications out of it; this is certainly a case that someone could make. However, the definition proposed here asks for an explicit declaration of the larger goals that the measurement is supposed to address, and though it is possible, it seems unlikely that someone would attempt to justify his or her practice by making that kind of claim. Even if someone were candid enough to justify it this way, we would immediately understand that the "measure" is at best of extremely limited interest and usefulness. Any expectations that we could have previously had regarding such a measurement are immediately tempered by the narrowness made explicit in the goal. But what about the case where a scholar simply wants to publish a paper with measurement to put some ideas "out there" so they can be part of the discussion? Would about that kind of utility? The question in this case would be, what is the role of the measurement claims in

the paper. If the measurement claims are not central, or perhaps even accessory, then we can ask why they were included, but ultimately will evaluate the value of the new ideas outside a discussion about measurement. If on the other hand, the measurement claims are central, then the ideas in the paper are relevant to a measurement and should be examined accordingly. This is in the spirit of ceasing to consider "measurement" as a label—or a blessing—that ensures ultimate knowledge; we need to ask of each measurement, following E. W. Adams (1966), the details of both why it was made and how it was made.

Second, and going back to the point of how the adoption of an implicit "goal of measurement" informs different definitions of measurement, Michell's judgement regarding the irrelevance of "moderately useful predictions of other event" is predicated on a perspective of science in general, and measurement in particular, where you either figure out *the way things really work* or not: Did you discover truth about reality or not? is it really measurement or not? That is certainly one approach to valuing the practice of science and measurement, alternatively, this PPM would promote recognizing that such "moderately useful predictions" are more or less valuable depending on the context of application. Granted, a scale of clinical depression constructed under a measurement by fiat stance may not reveal ultimate truth about the quantitative or qualitative status of "depression" and may not allow us to formulate laws that reveals the underlying mechanics of depression, but if it allows therapists to reduce the rate of suicides among their patients, I for one would not qualify it as a useless outcome or that it is not of interest to psychological science. I contend that a criterion so stringent that would consider this as useless to science is predicated on the complete disconnect between the role of science and measurement with the practical concerns of human beings, which albeit is a possible stance, is contrary to the perspective advanced here.

Another issue that may raise concerns about the idea of Pragmatic justification is that when it has been invoked in the past, it has been usually interpreted as allowing any kind of inference, past or future, about the attributes that are being measured. In this light Pragmatic justification seems cheap, its criterion weak and too permeable. However, I contend that this is product of a very specific interpretation, that I call the *"any use, any time"* approach to Pragmatic justification, which basically assumes that the mere possibility of a use is enough to support any use. The problem with

a *any use, any time* approach is that it seems just too lax; if taken to heart, mere possibility of some use, some time, under some unspecified context is justification enough: perhaps my anxiety questionnaire has no practical applications right now, but it might!

There is of course nothing wrong with entertaining potential future applications, I would simply argue that they should not be equated with actual applications, after all, many procedures could in the future be successfully used to classify, order or quantify under a certain context, but until that happens, claiming to have measured seems premature. In contrast to the *any use, any time* approach, Pragmatic justification can be understood to be emphasizing that the successful measurement procedures are connected to clearly specified uses under clearly specified contexts of application.

Take for instance Lord and Novick (1968) justification of the assumption of a continuous variable underlying the performance of respondents in their classic book *Statistical Theories of Mental Test Scores* saying:

> This could be considered as an arbitrary strengthening of our model. However, *from a pragmatic point of view* the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale. If we construct a test score by counting up correct responses (zero-one scoring) and treating the resulting scale scores as having interval properties the procedure may or may not produce a good predictor of some criterion. (p. 22)

This is in principle consistent with Pragmatic perspective presented here (although Lord and Novick appear to have been using the word *pragmatic* without a special commitment to the the philosophical school), it clearly anchors the measurement claim of any given instrument to a specific purpose for which we are supposed to have empirical evidence that validates it as, for example, a predictor.[26] However, if we accept the *"any use, any time"* interpretation, then "the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale" becomes a trivial constraint, after all, I can always use an educational test to start

[26] Although some, like Hand (2004) for example, insist that Pragmatic criteria are limited to prediction, there is in principle no reason for such limitation. A Pragmatic perspective can seek, for instance, manipulation or explanation.

a fire (any use) and I can always claim that is just not useful yet (any time). Although these examples may seem silly, they highlight that the problems with a Pragmatic justification emerge when we forget that if a measure was justified for a given purpose in a specific context of application, any measurement claim and the inferences derived from it are also constrained to that purpose. In the future, other potential uses can emerge, potentially expanding the scope of the original measurement claims, but such expansion must be justified by showing actual "cash value," not merely on the basis of promissory notes and good intentions.

From the Pragmatic Perspective of Measurement advanced here, *a Pragmatic justification is not a blank check on inference*. If we are to invoke Pragmatic justifications, we should take them seriously. A Pragmatic justification defines the goals we are trying to accomplish, guides our selection of the attributes of interest, informs the aspects of our models that are relevant, provides desiderata for evaluating whether our measurement is useful or not and, most importantly, anchors our inferences to the context of the goal that we have defined. Generic statements regarding unspecified practical applications aimed at justifying otherwise unwarranted inferences are simply poor Pragmatic justifications. This kind of poor justification is made worse when the "pragmatism" is restricted solely to the methods section, but the conclusions are written as if by a metaphysical realist.

What does a Pragmatic, contextualized, justification of a measurement looks like? We can find one in the history of intelligence measurement. Allow me to start with a brief digression on the work that Alfred Binet, usually considered the father of intelligence measurement and testing, had done on the topic of psychological measurement before working on the field of intelligence (cf. Michell, 2012a). While studying the possibility of measuring suggestibility through an experiment involving the judgement of line lengths, Binet (1900)[27] had explicitly differentiated between quantitative measurement and ordering:

> All physical measurement, when it is precise, gives not only an ordering of measured objects, but furthermore an indication of the number of times that one object is larger, heavier, etc. than another, that is, an

[27]There is no current English translation of this book. I thank Rebecca Freund for kindly translating the relevant passages discussed here.

> indication of the number of times that such a quantity contains the unit. This is not the same in psychological measurement; also, I think that this is not true measurement; it is quite simply ordering.[28] (pp. 103–104)

Why is it relevant that Binet made this distinction? This will become apparent presently, please bear with me. Years later, when presenting his effort (in collaboration with Simon) in the construction of an intelligence scale, Binet and Simon (1916) begin by stating in the first line of their article (in language common to their times):

> *Before explaining these methods let us recall exactly the conditions of the problem which we are attempting to solve* [emphasis added]. Our purpose is to be able to measure the intellectual capacity of a child who is brought to us in order to know whether he is normal or retarded. We should therefore, study his condition at the time and that only. We have nothing to do either with his past history or with his future...

It was in the context of making diagnostic decisions of students in primary school, and with that goal in mind, that Binet and Simon (1916) state[29]:

> This scale properly speaking does not permit the measure of the intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured, but are on the contrary, a classification, a hierarchy among diverse intelligences; *and for the necessities of practice this classification is equivalent to a measure* [emphasis added]. (Binet & Simon, 1916, pp. 40–41)

This is where the first quote about quantitative measurement and ordering becomes relevant. Although it did not make it to the English translation, Binet includes in the original French[30] a footnote to the word "measure," explicitly referencing his

[28] In the original French Binet uses the word *classement*, which can mean, depending of the context either classification or ordering, and the context of this passage seems to be used in this last sense.

[29] This quote was brought to my attention by the detailed analysis that Michell (2012a) makes of it and its relation to the concept of heterogeneous orders, although he draws markedly different conclusions from it.

[30] In the original French this quote also uses the word *classement*, but in this case the English translation treats it as "classification" instead of "ordering." In light of the omitted footnote and the overall context, it is more likely that Binet meant "ordering" instead.

previous treatment of quantitative and ordinal attributes cited above. Binet understood the distinction between quantification and ordination, and while he acknowledged it, he stated that "for the necessities of practice"—which he made explicit in the first line of the article—his scale can be used for measurement. Here Binet proposed an ordinal scale of intelligence to be used for diagnostic decisions of students in primary school in the early 20th century France, and provided evidence to support that use.

Binet was not measuring "Intelligence" as a universal variable, his focus was not the formulation of psychological laws, and he embarked in the construction of an ordinal measure to address a practical challenge. According to the Pragmatic criterion proposed in this paper, there is no such thing as an "Intelligence Measure" without a qualification of the context in which it is supposed to be useful. This raises the question of how to define a context, and what makes a context A similar enough to context B to justify the use of a measurement instrument developed for A applicable to B, or what makes context C different enough to make such a transfer questionable. Defining what is relevant or irrelevant in a context is not a trivial task, and arguments for or against of using a instrument in a new context will need to be examined in a case by case basis; then again, this kind of thoughtful consideration is already advocated for instance in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014).

If like Binet, we are engaged in solving a specific diagnostic or selection issue, our evaluation of that measurement should be according to that purpose, and the extent of our inferences should be also constrained to that purpose. If psychologists want to conceptualize and measure intelligence as an attribute akin to temperature, and/or as part of a set of psychological laws, certainly an ambitious goal, they will need to go beyond the same evidence that Binet used to justify his measure, showing that "Intelligence" does behaves in law-like relations—just like temperature does–with other attributes that have been quantified. As Sherry (2011) so clearly states:

> Acceptance of the pragmatic approach to quantification does not, of course, entail the existence of quantitative psychological attributes. *The pragmatic value must be demonstrated* [emphasis added]. The conceptual advance that we admired in temperature measurement has no obvious

> analogue in, for instance, intelligence measurement. The general intelligence factor, g, does not occur in psychological laws in the straightforward way that temperature occurs in $m_1c_1(t_1 - t_f) = m_2c_2(t_f - t_2)$ and $\Delta Q = mc\Delta t$. g correlates mathematically with academic success, income, etc., but it is not clear that the correlation provides greater understanding when g is treated as cardinal rather than ordinal data.

Before this kind of evidence is brought to bear, our measures of intelligence are tailored, and anchored, to more modest goals. To those interested in creating measures for establishing quantitative psychological or social laws, I wish them well—and borrowing a phrase from Duncan (1984, p. 170)—"I shall be interested in seeing the results but will only insist that a program not be mistaken for the accomplishment."

2.6 Contrasting and Comparing

How does this proposed Pragmatic definition compare to the other definitions explored in section 2.3? I will approach this question by looking into the four components of this definition, namely, the goal of measurement, the activities involved in it, the nature of the attribute, and the model of the attribute.

2.6.1 The Goals of Measurement

Perhaps the clearest difference from the Pragmatic definition versus the alternative definitions reviewed so far is the explicit inclusion of the goals of measurement as part of the definition. Neither the main theories—classical, operational, representational—nor the two application-oriented approaches—latent variables and metrology—discuss goals as an integral part of the definition of measurement.

As pointed out in section 2.5.4, although the goals of measurement are not discussed as part of these definitions, measurement scholars often discuss what they consider to be the goals of measurement as part of the background to their definitions. In this sense, the unstated goals shape the definitions of measurement of different approaches.

In the case of Michell and the CTM, the goals of measurement, and science more generally, fall clearly within the tradition of the spectator theory of knowledge, with

science aiming to "discover nature's way of working" (Michell, 1990, p. 27), and measurement aiming to "get to know about *quantitative* attributes" (Michell, 2008c, p. 137). This emphasis on measurement as portrayal is shared by the RMT tradition, where the focus is "how numbers enter into science" (Narens & Luce, 1986, p. 166), attempting on the one hand "to understand the nature of empirical observations that can be usefully recoded, in some reasonably unique fashion, in terms of familiar mathematical structures" (Luce & Suppes, 2002, p. 1), while on the other hand attempting to identify laws—in the same way as Campbell (1920)—based on the application of these mathematical structures (Luce & Suppes, 2002; Narens & Luce, 1986).

Again, notice that all these discussions about the goals of measurement, and the relation to the goals of science are not explicitly part of the definitions of measurement either in CMT and RMT. These discussions take place in introductions and backgrounds, and in these cases focus on *the* goal of measurement, shaping the respective measurement definitions and theories in ways that completely ignore different goals of measurements as a relevant aspect.

Adams' (1966) criticism of RMT for ignoring what measurements are made *for* is equally applicable to the CTM, as both of these theories appear to be solely concerned with what measurement ought to be in order to adequately create theories within research contexts, while disregarding practical ones. In other words, CMT and RMT as theories are shaped by their assumption that there can only be one goal for any and all measurements, namely, the discovery of reality's laws.

For its part, measurement under operationalism as advocated by Dingle (1950) was ultimately concerned with the establishment of correlations "so as to afford a true picture of the relations which the observations exhibit" (p. 5). For Dingle, and presumably for Bridgman too, these correlations would no longer be simply a function of the external world, as they would depend on the operations performed to investigate them. However, despite this departure from the focus on the accurate portrayal of the world in CMT and RMT, operationalism simply transfers the idea of portrayal from the attempt to capture a single reality to the attempt to accurately portray the correlation of observations that emerge as a function of certain operations. In this version of operationalism, there is again a single overarching goal (which is much more general than in the previous two theories) that all measurements are after: "the process of measurement is specified only in those details which are necessary to

ensure that the results obtained are useful for our ultimate purpose of correlation" (Dingle, 1950, p. 14). Although in Dingle's operationalism the definition of measurement is so broad that it effectively it barely tells us what measurement should be, it stills frames measurement as a single purpose activity, in this case the very broad search for correlations.

In the case of the two application oriented approaches, what measurements are made *for* is still not explicit in their definitions, but it is discussed actively in the surrounding context.

Given its close ties with statistical modeling, within the LVM tradition, the goal of measurement can often be conceived as the search for "good fit" between the statistical model and the data. However, this is by no means a unanimous perspective among those who rely on latent variable models in measurement contexts (Mislevy & Haertel, 2006; Rasch, 1960/1980; Wilson, 2005; Wright, 1997). Moreover, the professional communities that rely on these models to serve the public and inform policy—in addition to conducting research—have actively tackled the issue of the intended use of measurements. This concern hit the mainstream psychological and educational assessment community in the 1980's with the publication of the 1989 chapter on validity by Samuel Messick (1989) in the book *Educational Measurement*. This publication strongly emphasized a perspective where:

> ...the key issues of test validity are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of the social consequences of their use. (Messick, 1989, p. 13)

This emphasis on the intended use and the consequences of measurement was later on incorporated into the 1999 *Standards for Educational and Psychological Testing*, a consensus document generated by three major professional associations: The American Educational Research Association (AERA), The American Psychological Association (APA), and the National Council for Measurement in Education (NCME).

In the face of the recognition that this perspective has achieved, it seems surprising to have previously declared that in the realm of latent variable modeling in particular, or in educational and psychological assessment in general, there is no explicit consideration of the goals—what measurements are for—in this approach. The

reason for this seeming incongruence is that these two documents refer explicitly to the intended use of *tests and test scores*, which is used more or less interchangeably with measurement. As Maul (2014) indicates in his review of the 2014 version of the *Standards*:

> The term "measurement" is used throughout the Standards, but it is never defined; it appears to be used interchangeably with the terms "testing" and "assessment." Such an omission and identification could only be considered defensible if the concept of measurement had been thoroughly established to be synonymous with that of testing, but this is patently not the case, as has been pointed out repeatedly by others (e.g., Borsboom, 2009; Hood, 2009; Michell, 1999).

The ambiguous, and potentially misleading, identification of testing and measurement is a symptom of the larger disconnect between practitioners of educational assessment and the larger measurement literature, such that measurement as a concept is defined separately (usually following Stevens' 1946 definition) from the concept of validity, in which the role of the intended use has been incorporated.

Finally, in the case of metrology, the goals of measurement are referenced in one of the three notes that complement the definition of measurement in the VIM (Joint Committee for Guides in Metrology, 2012), where it is stated that "measurement presupposes a description of the quantity commensurate with *the intended use of a measurement result*, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions." [emphasis added] (p. 16). In this way, what measurements are made for (i.e. "the intended use of a measurement result") is dealt with in the same level as the other more engineering aspects of measurement.

The differences between these theories and this PPM regarding the goals of measurement are summarized in Table 2.2, indicating the role that goals play in the theory, what is the goals that are considered, and whether measurement is conceived to be a single or multiple goal activity.

Table 2.2 highlights that an important difference between this PPM is that the goals of measurement play a central role, being explicitly incorporated as part of the

Table 2.2: Comparison Between Theories - Goals of Measurement

	CTM	OPER.	RMT	LVM	METR.	PPM
Role	*Implicit*	*Implicit*	*Implicit*	*Acknowledged*	*Acknowledged*	*Central*
Goal	*Discovery*	*Correlation*	*Find Laws*	*Intended use*	*Intended use*	*Intended use*
Scope	*Single*	*Single*	*Single*	*Multiple*	*Multiple*	*Multiple*

proposed definition of measurement. This contrasts is stark against the implicit influence that the goals of measurement have in CTM, operationalism, and RMT, where the authors do not consider them as part of their definition of measurement. Within application-oriented theories, on the other hand, the role of measurement goals has been acknowledged to different extents, but it has not been incorporated as a central part of the measurement definition, with Metrology almost doing so by acknowledging the role of intended use in one of the notes that accompany the VIM definition.

2.6.2 The Activities of Measurement

As pointed out in Section 2.5.1, many if not all of the definitions of measurement discussed in this paper describe it as some kind of activity. The central point that differentiates the proposed PPM is the emphasis on what I contend is a more useful level of description of the activities that we engage in when we want to measure, namely, quantification, ordination and classification. This level of activities contrasts with the characterization in the other theories, such as *numerical estimation* in the CTM (Michell, 1997b) or *numerical assignment* both in the RTM (Campbell, 1920; Stevens, 1946) and LVM, as well as *numerical production*—as a function of a set of operations—in operationalism (Dingle, 1950). The closest definition in terms of the level of description of the activities is perhaps the one offered by Metrology, where the activities that are indicated in the definition are correspond to "obtaining and reasonably attributing quantity values" (Joint Committee for Guides in Metrology, 2012).

Despite the difference in the level of description of the activity, it is possible to examine which of the three activities that are explicitly identified in this PPM are consistent with the different theories reviewed in this paper. A comparison regarding the activities of measurement between these theories and the PPM presented in this paper is presented in Table 2.3.

In this sense, the most restrictive perspective is advocated by the CTM, where it is clear that only one of them—quantification—is rightfully considered as measurement. Both ordination and classification deal with qualitative distinctions and are better left to be examined through other—non-measurement related—means. All the other theories of measurement here reviewed adopt more expansive definitions, including at least ordination as part of measurement.

Among the theories that favor the inclusion of both quantification and ordination, but leave out classification, we can find the metrological definition found on the VIM (Joint Committee for Guides in Metrology, 2012), where they strike an interesting middle ground: only quantities can be measured, but they include "ordinal quantities" in that group. Although none other of the main theories reviewed in this paper subscribe just to the exclusion of classification as a form of measurement, this position can be found among other measurement scholars, being subscribed for example by Torgerson (1958), Berka (1983), and Duncan (1984).

Table 2.3: Comparison Between Theories - Activities of Measurement

	CTM	OPER.	RMT	LVM	METR.	PPM
Activity	*Numerical estimation*	*Numerical production*	*Numerical assignment*	*Numerical assignment*	*Obtain and attribute quantity values*	*Quantification, ordination, and classification*
Quantification	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Ordination	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classification	*No*	*Yes*	*Yes*	*Yes*	*No*	*Yes*

The remainder of the theories reviewed in this paper would more or less explicitly accommodate classification, ordination and quantification as part of measurement. Operationalism, at least as discussed by Dingle (1950), recognizes that some operation will generate a number, but it does not prescribe or limit them in any way, which in principle would allow quantification, ordination and classification (among other possible activities) as long as they are conceived as a set of operations. The representational measurement theory, ranging from Stevens (1946) to Narens and Luce (1986), through its focus on scales of measurement, effectively encompasses all three of the activities too. However, it is important to note that RMT is not really predicated over the distinction of activities at the level of quantification, ordination and classification,

but on groups of mathematical transformation, which define many possible scales, including some that could arguably be located in between them, for example, the hyper-ordinal scale (Suppes & Zinnes, 1962). It is worth mentioning though that classification, associated with nominal scales in RMT is hardly discussed in this literature (they are barely mentioned once by Krantz et al. in the first volume of *Foundations of Measurement* for instance), presumably because they are of less interest mathematically speaking. Finally, Latent Variable Models in general encompass statistical models that support all three of these activities, but are not limited to them.

2.6.3 The Attributes We Measure

The nature of the attributes that are measured is another facet that differentiates the proposed PPM from the other theories discussed here. To illustrate this contrast, consider that the CMT—as proposed by Michell—construes measurement as the estimation of ratios of magnitudes which are part of reality, independent of any context or operations used to measure, while Operationalism—as proposed by Dingle—abandons the assumption of an underlying attribute and focus exclusively on the operations that are used to measure. Put a different way, for CTM the ratios that we want to estimate exist independent of what we do, while for Operationalism, the attributes that we want to measure are a product of what we do.

From the theories that were reviewed in Section 2.3, CMT has an explicit realist position about the attributes that are measured, having something to say about their ontology. In addition to CTM, and although not explicitly committed to realism, the language used in Metrology to define quantities seems clearly realist: "property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference" (Joint Committee for Guides in Metrology, 2012, p. 2). This is not surprising considering that the Metrological community has been traditionally associated with the physical sciences, where realism seems to be the default position.

For their part, both Operationalism and RMT are different flavors of empiricism, and as such, they either stay away from ontological discussions in the case of RMT, or completely disavow them, as is the case of Operationalism. For RMT, all we have is qualitative data that we observe, and mathematical structures that we impose on

them, without any discussion about the ultimate reality (or lack-thereof) of those observations. Operationalism for its part was conceived by Bridgman as a way of inoculating science from unstated ontological assumptions that could lead it astray, and therefore treated any talk about the underlying properties merely as a time-saving convenience at best. It is worth mentioning however that although these theories in their pure form may be incompatible with realism, that does not mean that they are always treated separately in practice. The mathematical developments of RMT have been accepted and incorporated as part of realist theories, as is the case for Michell (1990, 1999, 2005) and others. As Tal (2015) points out:

> While their stance towards operationalism and conventionalism is largely critical, realists are more charitable in their assessment of mathematical theories of measurement. Brent Mundy (1987) and Chris Swoyer (1987) both accept the axiomatic treatment of measurement scales, but object to the empiricist interpretation given to the axioms by prominent measurement theorists like Campbell (1920) and Ernest Nagel (1931; Cohen and Nagel 1934: Ch. 15).

At the same time, Operationalism has been adopted in social sciences, and particularly psychology, through the idea of "operational definition" while ignoring its philosophical background, such that it can be used in methodology sections that support realist conclusions.

The LVM tradition is perhaps the most clearly agnostic perspective regarding the nature of the attributes that are measured, where the label *construct* can be used to refer to diametrically opposed interpretations of the models used for measurement.

On one extreme, latent variable models can be understood as methods of data reduction, where we are simply reducing the complexity of multiple observed variables by calculating a more manageable summary, the latent variable. From this perspective, the latent variable does not represent an underlying attribute present in the world, but is simply a convenient summary. On the other extreme, latent variable models are to be understood from a realist perspective (Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2003):

> Entity realism about latent variables is needed to motivate the choice for the reflective model over the formative model. Theory realism follows from the observation that the formal side of the theory implies that it is possible to be wrong about the position of a subject on the latent variable, and that weaker formulations—using empirical adequacy instead of truth—are difficult to interpret. (p. 216)

Table 2.4: Comparison Between Theories - Attributes

	CTM	OPER.	RMT	LVM	METR.	PPM
Attributes	*Quantities*	*Results of operations*	*Observed qualitative data*	*Constructs*	*Quantities and ordinal quantities*	*Relevant attributes*

2.6.4 The Model of the Attribute

As indicated in Section 2.5.2, in this paper the model of the attribute refers to the theory that defines which objects can be conceived as having the attribute, establishes what kind of relations characterize the attribute, guides our expectations regarding potential study and manipulation, and provides the basis for justifying why an instrument is an appropriate indicator of the attribute of interest. Although all the measurement theories discussed here recognize the role theories play in some or all of these aspects, they do not all agree on the prominence that these models play in their measurement definitions or the role that these theories have in the process of measurement.

A first element to note regarding the role of models is that both CMT and RMT have been traditionally uninterested in the details of how the theoretical models interact with the engineering aspects of measurement (i.e. the *instrumental task* of measurement for Michell, as opposed to the *scientific task*). As Tal (2015) points out when discussing the process of instrument calibration "traditional philosophical accounts such as mathematical theories of measurement do not elaborate on the assumptions, inference patterns, evidential grounds or success criteria associated with such methods." As such, these two approaches favor an abstract approach to measurement. This tendency of CMT and RMT contrasts with the opposite emphasis in operationalism and

the approaches in the applied realm, where there seems to be a great deal of emphasis regarding the instrumental aspects of measurement. Although the introduction of this PPM has focused on the more philosophical aspects of measurement, the instrumental aspects are acknowledged as part of the model of the attribute, the selection of the procedure for quantification, ordination, and classification, as well as the goals of measurement. In this sense, this PPM follows E. W. Adams that "the two basic things to know about this measure are: (1) what it is used for, and (2) how it is made" (E. W. Adams, 1966, p. 133).

Table 2.5 summarizes the differences between these theories and the PPM presented here in terms of their conceptualization of models. both in terms of the level at which the models are specified and the role that the models play in them.

Table 2.5: Comparison Between Theories - Models

	CTM	OPER.	RMT	LVM	METR.	PPM
Level	*Abstract*	*Instrumental*	*Abstract*	*Instrumental*	*Mixed*	*Mixed*
Models	*Descriptions*	*Eclectic*	*Descriptions*	*Eclectic*	*Eclectic*	*Tools*

The focus on the instrumental aspects is a characteristic of Operationalism, where the operations involved in the measurement process constitute in themselves the attribute that is measured. For their part, the VIM in the field of Metrology and in the *Standards* in the fields of educational and psychological measurement also cover the instrumental aspects, with the *Standards* focusing solely on them and the VIM and GUM covering both the conceptual as well as instrumental aspects.

A second element when contrasting the role of models between these approaches is the role that the model is playing. For CTM, theories are attempts at descriptions of reality, and are usually not discussed in terms of models by Michell (1999, 2005). A similar role of description is played by substantive theory in RMT, where the aim is not to achieve an accurate description of an underlying reality, but to faithfully represent the relations observed in data. To the extent that CTM or RMT would consider that models play a role in measurement, these models would be considered descriptions, the former of reality, and the latter of observed data. This emphasis in *models as descriptions* is in contrast with an emphasis of *models as tools*, that while is not explicitly adopted in Operationalism, LVM or Metrology, is potentially consistent with them.

In this sense, the PPM adopts the *models as tools* perspective, in accordance with the anti-representationalism characteristic of Pragmatism.

2.7 Summary

I have advanced a proposal for extending Pragmatic ideas into the field of measurement, these ideas are organized in relation to a Pragmatic definition of measurement presented in Section 2.5:

> Measurement is an activity of classification, ordination, or quantification of a set of elements according to a model of a relevant attribute in service of a larger goal.

In keeping with the Pragmatic ideas presented in Section 2.2, this definition is provided as a tool. This particular tool is aimed at gauging measurement claims in terms of their usefulness in the face of the diversity of practices that are conducted under the banner of "measurement."

At the same time I cannot stress enough that this definition, and the set of questions that it attempts to raise, are not—nor pretend to be—exhaustive. One major issue, the specifics of the how and all the details of instrumentation, merits its own investigation, especially considering that each prototypical conception of measurement will likely carry its own standards for analyzing measurement methodologies.

Having said that, this PPM is not solely restricted to the criterion of utility, but also presents a Pragmatic account of the attributes that we try to measure. I advocate that it is in our best interest to consider attempts to classify, order and quantify as prominent activities, in service of the different models that we can develop in our attempts to study, predict and manipulate the attributes that we consider relevant.

Going back to the imaginary dialog presented at the very beginning of this paper, when presented with a measurement claim such as "I measured anxiety?," the PPM invites the skeptic to stop asking "did the researcher *really* measure anxiety?," and ask instead:

- What were you doing? Classifying, Ordering or Quantifying?
- What is the attribute that you were measuring?

- What is the model that you have of that attribute?

and last, but certainly not least:

- For what purpose are you measuring this attribute?

These are questions that can provide us with a general understanding of what is behind that measurement claim, clarifying what the attribute is supposed to be and what are the criteria for success according to the purpose of the measurement. While some may argue that this is too broad a definition, I would respond, that this definition is tailored to those that are more interested in understanding the what, the how and the why of the practices commonly referred to as measurement, than those whose focuse is in elucidating whether each one of them *really* is or not measurement.

It is my hope that this PPM contributes to a better understanding between researchers, practitioners and users of measurement, by bringing to the foreground an understanding of measurement where measures are not always concerned with quantification, that attributes need not be considered as natural kinds or universals, that understanding the model of the attribute that underlies a measurement is central to designing and interpreting measurements, and that the purpose for which the measurement was developed informs us about the scope of its utility, both to judge its success as well as limiting the inferences that can be supported by it.

Chapter 3

Categorization, Ordering and Quantification: Selecting a Latent Variable Model by Comparing Latent Structures

David Torres Irribarra & Ronli Diakow

3.1 Introduction

The previous paper proposed a Pragmatic Perspective of Measurement (PPM), from which measurement is understood as "an activity of classification, ordination, or quantification of a set of elements according to a model of a relevant attribute in service of a larger goal." While that paper focused on a theoretical level to canvas alternative measurement theories and put forward a Pragmatic perspective as an alternative, this paper delves into a more practical level, specifically, the discussion of how *the model of a relevant attribute* is formalized in the statistical framework used for the analysis, and specifically, how these different models can be used to test whether it is reasonable, from a statistical perspective, to adopt a model of classification, ordination, or quantification in order to characterize a relevant attribute under study.

The use of statistical models in the social sciences to make inferences about attributes is widespread. The adoption of a statistical model is often done as a matter

of convention within a discipline, such that the same data might be analyzed using analysis of variance in psychology, item response theory in education, or latent class analysis in sociology. However, the selection of these models is not trivial, as alternative statistical models make different assumptions both about the structure of the observed data as well as the latent structure of the attribute that is being investigated. Although it might be the case that in certain cases models with different assumptions yield results that are consistent in practice, it seems unwise to ignore the possibility that different models could potentially lead us to very different conclusions regarding the way in which we conceive of a relevant attribute and how we decide to act on this conception.

The aforementioned PPM highlights three kinds of activities: classification, ordering, and quantification. In pursuing one of these activities through the use of statistical models, the researcher or practitioner can select from a vast pool of possible models in order to accomplish their objective.

Ideally, we want consistency between our theoretical models of a relevant attribute and the statistical models we use to make inferences that attribute, at least in order to facilitate the interpretation of results and the inferences we make based on them. From this perspective, if our theory characterizes an attribute as a set of qualitatively distinct classes, we would seek to adopt a statistical model within the pool of models that would more or less reflect such structure. Similarly, if our theory describes the relevant attribute as a quantity, we would then seek to adopt a model that would reflect that kind of structure.

This process would most likely involve the comparison of somewhat similar models with certain variations—one or more additional covariates, one or more additional random effects, correlated or uncorrelated effects, etc.—and a subsequent process of selection of the best fitting one through some statistical procedure. However, following such a procedure is predicated on accepting the theoretical characterization of the latent structure of the relevant attribute without questioning. In other words, while we usually endeavor to use statistical criteria to select the most appropriate model within each of these activities, this logic is not often applied to the selection of models across them.

Beyond the expectation that our models should to some extent be coherent with the structure of the attributes described in our theories, it is critical that the inferences

we make about the attributes can be reasonably justified by the statistical models used to generate them. Researchers should question the assumptions that serve as the basis of their statistical models in order to inform their hypotheses about the structure of the differences in a relevant attribute, and as a way of gathering evidence that the structure assumed by a statistical model can reasonably support inferences based on that such assumption. This paper presents a framework for actively examining this issue. Through the use of a model selection framework based on latent structure analysis, which relies in successive model comparisons, it is possible to evaluate whether it is warranted, from a statistical perspective, to use a model that characterizes the latent structure as one of differences in kind, order or quantity.

3.1.1 Tenable Assessment

The unexamined acceptance of the structure of an attribute of interest implies the acceptance of a specific way of construing——and making inferences about—the elements that are being studied, which in the social sciences are often persons. Stated more concretely, accepting that intelligence or anxiety is a continuous quantitative variables at the moment of modeling persons' responses to a test will lead to inferences that will cast interindividual differences in those attributes as ones defined by a single scalar, and simply accepting that addiction patterns or reading habits are qualitatively distinct when modeling data from a questionnaire will cast interindividual differences in those attributes in terms of unordered classes. From a PPM it can be argued that adopting these assumptions might be reasonable under some circumstances, but such justification requires the examination of the issue, and most likely will require the comparison of that model with alternative ones. When we quantify interindividual differences, we claim to be able to characterize differences in the relevant attribute as distances; when we create a rank of such differences, we claim to be able to determine relations of more or less; finally, when we classify interindividual differences, we claim to be able to differentiate qualitatively different kinds.

It is important to note that we can make several distinctions at different points regarding the structure of the attribute under a PPM. At a general level, we may be quantifying, ordering or classifying, which speaks to the kinds of inferences we intend to make to address the broader goal of inquiry motivating a measurement

procedure. We can also have a theory which characterizes the relevant attribute as having a specific structure; this structure can be more specific, say, defining whether it is a discrete or continuous quantity, or whether the different classes may be fully ordered or partially ordered. Finally, we can decide to model the attribute under a different structure, having the flexibility of choosing a different model from the one defined by our theory. For instance, I may be interested in ordering teachers according to a theory that indicates that there are four ordered teacher performance levels, but I may choose to use a quantitative model when analyzing data from a teacher portfolio, which I will later will convert into a set of four categories. This is, for instance, a common way of going about conducting assessments that aim to classify teachers.

There are many possible ways of conceptualizing the structure of a variable theoretically, and also many possible ways of modeling such structures mathematically in ways that may or may not coincide with the theory. Focusing on the latter, Figure 3.1 illustrates five different structures (or types of structures) that could be considered when modeling interindividual differences on a variable of interest depending on the theorized underlying structure of the variable. These types of structures—as there are multiple ways of defining each one of them[1]—are only a subset of the many different ways in which we can model latent structures, and the selection is organized specifically in terms of expressing variations between a quantitative structure and a classificatory one; moreover, these are but a sample of the gamut of possible models that could be conceived within that range. These structures may range from a single continuous quantitative distribution, illustrated in panel (a), or a mixture of continuous distributions, as in panel (b), to a single set of latent classes, either located in a continuum (c), ordered (d), or qualitatively different (e).

When faced with a dataset, a key question is which of these structures offers a better account of the data—usually in the context of statistics this is determined in terms of some fit metric—allowing us to (i) critically question whether the structure inferred from the data is consistent or not with our theoretical assumptions, and (ii) consider whether the structure of the data lends support to inferences that assume one structure or the other.

[1] For instance, the models in pane (a) can potentially include the Rasch Model (Rasch, 1960/1980) or the 2PL and 3PL models (Birnbaum, 1968)

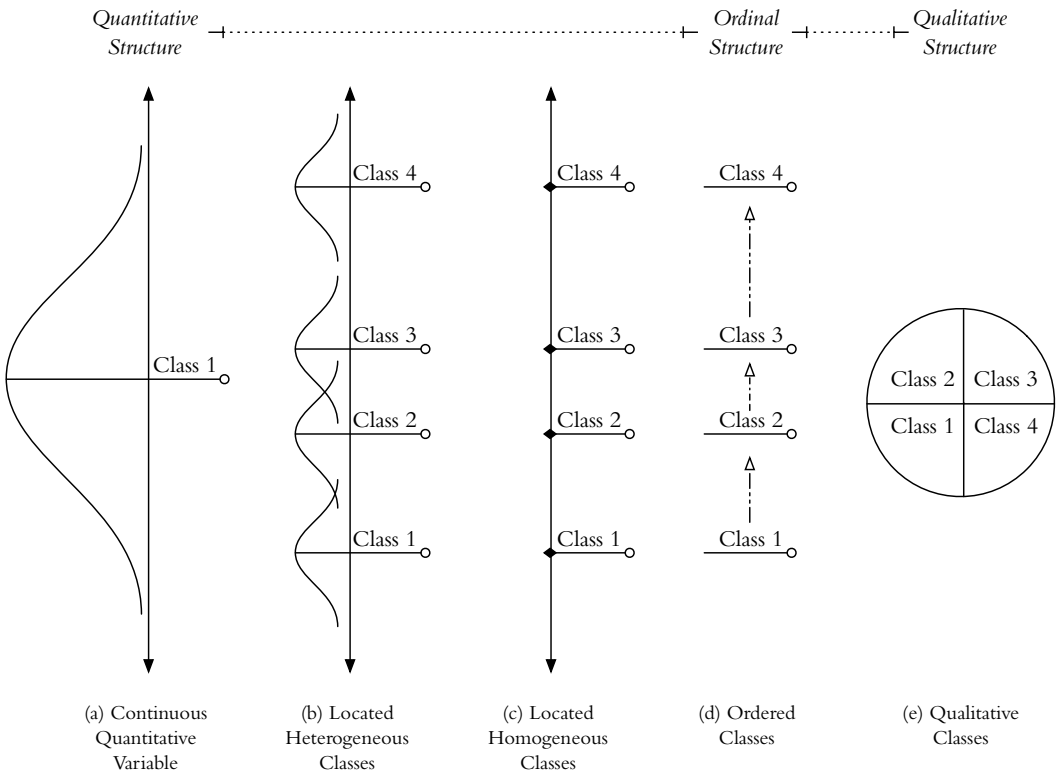


Figure 3.1: Five of the Possible Types of Model Structures That Can Be Attributed to a Latent Variable.

We contend that addressing these issues requires the examination of the statistical assumptions that are being made regarding the structure of the attribute of interest. The proposed framework organizes some basic latent variable models according to two kinds of restrictions placed on the latent variable, namely:

Monotonicity, which implies that the probability of responding correctly is non-decreasing as a function of increasing proficiency and/or that is non-increasing as a function of increasing difficulty.

Scale, which implies that the differences between persons (and items) are quantitative in nature.

We present a method of model selection that tests the empirical support of these assumptions based on comparisons among these latent structures.

3.1.2 Latent Structure Analysis

This paper will rely on the framework of latent structure analysis (LSA; Lazarsfeld & Henry, 1968). We have selected LSA for its generality and flexibility. Latent structure analysis allows us to formulate all the models that we will be using under a single framework. Additionally, specific cases of LSA, such as latent class models, item response models, and factor analysis, represent the dominant approaches for data analysis in a variety of disciplines. Models that fall under the umbrella of LSA, and the discussion of these models, should be both familiar and applicable to a general social science audience. By presenting our ideas in a familiar framework, we hope to provide an entry point for social science researchers towards modifying or extending their current data analysis practices.

3.1.3 Overview of the Paper

The purpose of this paper is to introduce a framework for selecting statistical models that supports assessment inferences by explicitly testing assumptions about the structure of the variables.

To this end, the second section of this paper presents a more complete discussion of latent variable models under the framework of latent structure analysis . This includes a historical overview and a review of the statistical models. We then present our ideas regarding the model selection framework in more detail in section 3.3, where we discuss our model framework and theory for model selection. The paper concludes with a discussion of the implications of this framework from the perspective of the PPM.

3.2 Latent Structure Models

Latent structure analysis (Lazarsfeld & Henry, 1968) is a data analysis approach that introduces latent variables to account for the observed pattern of associations between manifest variables (Heinen, 1993). According to Everitt (1984), the "latent" component refers to one or more hypothetical variables that are included in the analysis for the purpose of understanding some area of research, and for which there are no operational methods for direct measurement.

Latent structure models (LSM) allow us to summarize, understand, and predict the data patterns observed in the manifest variables in terms of the proposed latent variable. This is achieved through a mathematical model that relates the observed data (i.e. the manifest variables) to a theoretical construct (i.e. the latent variable). Latent structure models are probabilistic models that relate the probability of observed responses to underlying variables (Masters, 1985), many of which have similar functional form or can be expressed with similar functional form (Andersen, 1988). In terms of the mathematical models, latent variables can be defined more broadly as "random variable whose realizations are hidden from us as opposed to manifest variables where these realizations are observed" (Skrondal & Rabe-Hesketh, 2004, p. 1).

3.2.1 Local Independence & Dimensionality in LSM

The axiom of *local independence* is the "essential unifying characteristic" of latent structure models (Lazarsfeld & Henry, 1968, p. 22). This assumption is the commonality across latent structure models and the foundation from which (almost) all latent structure models can be derived (Clogg, 1988, 1995).

The assumption of local independence states that the observed responses are independent of each other conditional on the latent variable. This assumption is not only the cornerstone of the mathematical models of latent structure, but also the foundation of the their explanatory power (De Boeck & Wilson, 2004), which is derived from their capacity to account for the relationship between the manifest variables. Local independence can be partially relaxed via carefully specified models to account for violations under certain situations (e.g. defining models that account for correlated errors), but is necessarily implicit in all latent structure analysis to some degree (Clogg, 1995).

A second assumption that is commonly attributed to latent structure analysis, in particular in the context of item response models, is the assumption of unidimensionality. Though the assumptions of unidimensionality and local independence have the same implications for the structure and interpretation of the probability model, they are conceptually distinct. In practice, one of these assumptions could be violated even if the other is not.

Although it is true that any latent structure model indeed must rely on the assumption that the number of latent variables that account for the the observed relations is

correctly specified, this need not be a single latent variable. Several models have been developed to handle different kinds of multidimensional latent structures, for example confirmatory covariance structure models (Bollen, 1989) or multidimensional item response models (Reckase, 1997, 2009), which includes models such as the compensatory multidimensional random coefficient multinomial logit model (MRCML; R. J. Adams, Wilson, & Wang, 1997) and the non-compensatory general multicomponent latent trait model (GLTM; Embretson, 1984).

As McCutcheon (1987) points out, latent class analysis (LCA) models in general can also address multidimensionality underlying observed variables by formulating it in terms of typological classifications. Moreover, extensions in LCA have been developed to directly address multidimensionality, for instance, the multiple classification latent class model (MCLCM; Maris, 1999), which contributed to the development of another popular approach to multidimencionality within LCA framework, cognitive diagnostic modeling (CDM; Rupp, Templin, & Henson, 2010), and mixture item response models (Rost, 1990; Wilson, 1989) which connect the literatures of LCA and IRT. In the light of these different modeling alternatives it seems clear that the assumption of unidimensionality is not a characteristic necessarily shared by all latent structure models.

The challenges of identifying the correct dimensionality of an instrument is no trivial task, and extensive research has been dedicated to this issue (Junker, 1991; Stout, 1987, 2002), including recent work by Van Onna (2004) that directly explores the detection and impact of misspecification of dimensionality in ordered latent classes and non-parametric item response models. Although we recognize the importance of the issues related to appropriate specification of multidimensionality, in the present paper we will adopt unidimensionality as a simplifying assumptions in order to focus the scope of the discussion.

3.2.2 A historical overview of LSM

Factor Analysis

The development of latent structure models can be traced to the early 20th century in psychology in the factor analysis tradition originating with the influential—and ambitiously titled—work of Charles Spearman (1904): *"General intelligence": Objectively determined and measured*. Spearman's research in factor analysis pioneered the

assumption of a single continuous latent variable underlying the observed responses, sowing the seeds that would later lead to the developments of the factor analytical tradition (Gorsuch, 2003; Lawley & Maxwell, 1971; Thurstone, 1947), up to the work in confirmatory factor analysis (CFA) within the much broader framework of covariance structure analysis (Jöreskog, 1971), also known as structural equation models (SEM; Bollen, 1989).

It is important to note that, following the work of Spearman, the development of this framework throughout the 20th century considered both the latent and manifest variables as continuous. This assumption set the factor analytic tradition apart from the work from the developments on latent structure analysis, which are concerned with the analysis of data with categorical manifest variables.

Latent Classes and Latent Continua

Major advances in the field took place in the 1950's and 1960's. Two traditions arose for latent structure modeling that differed on whether the latent variables of interest were conceived of as categorical or continuous. Categorical latent variables models were explored under the banner of latent class analysis while continuous latent variables were studied under the name of latent trait models. These two traditions shared a focus on the analysis of categorical manifest variables, both dichotomous and polytomous, which set them apart from the factor analytic tradition mentioned above.

In sociology, the work of Paul Lazarsfeld and Neil Henry laid the foundations of what would become the latent class analysis tradition. In their 1968 book *Latent Structure Analysis*, Lazarsfeld and Henry coined the term and presented a framework that originally comprised continuous and categorical latent variables. However, subsequent work in this tradition would mainly focus on categorical latent variables and the use of latent class analysis for their study, with important contributions by Leo Goodman and Clifford Clogg (Clogg, 1979; Clogg & Goodman, 1984; L. Goodman, 1974).

In parallel to the developments by Lazarsfeld and Henry in sociology, an alternative tradition emerged that focused on the use of continuous latent variables, originating the domain known as item response theory (IRT) within the field of psychometrics. This tradition was based on the pioneering works of Ledyard Tucker (1946) and Louis

Guttman (1950), and was further developed by the work of Georg Rasch (1980), and the publication of Frederic Lord and Melvin Novick's *Statistical theories of mental test scores* (1968), a renowned work both on Classical Test Theory and also Item Response Theory due to the inclusion of a set of chapters by Allan Birnbaum (1968) on latent trait models. At the same time that this work was being developed, the field was also being extended into into Non-Parametric Item Responses Models (NIRT) with the work of Robert Mokken (1971).

Connecting latent classes and latent continua

A common framework for latent class and latent trait models extends back to Lazarsfeld and Henry (1968). Subsequently, a number of other authors have considered the relationship between them, including Bartholomew (1987), Langeheine and Rost (1988), Heinen (1996), Rost and Langeheine (1997), Skrondal and Rabe-Hesketh (2004) and De Boeck, Wilson, and Acton (2005).

The primary unifying feature of latent class and latent trait models is the idea of a latent variable. Lazarsfeld and Henry (1968) classified both of these under the umbrella of latent structure models, which they define as any model that proposes a latent variable to explain dependence in observed responses. However, as discussed above, much subsequent work distinguished whether the latent variable was discrete or continuous. Another distinction commonly made is that latent trait models rely on parametric functional forms while latent class models are not. Latent trait models define the relationship between the latent variable and the responses, the item characteristic curve, as a function of a set of parameters, while latent class models estimate the (conditional) probabilities directly (Masters, 1985).

One attempt to reconcile data analysis under the two traditions interprets latent class models as approximations to a latent trait model with latent variable forced to take only a few values, an approach that is also known as located latent class (Lazarsfeld & Henry, 1968; Uebersax, 1993). For example, Rost (1988), building upon the idea of located latent classes from Lazarsfeld and Henry (1968), links latent class and latent trait by considering a discrete distribution for the quantitative latent variable with groups along the latent variable. In terms of Figure 3.1, this approach uses the structure of model (c) to link models (a) and (e). This reflects the need to consider

scores for latent classes to coordinate between latent class and latent trait models (Clogg, 1988).

Although linking latent class and latent trait models in this fashion can be done both within the tradition of the two-parameter logistic model as shown by Uebersax (1993), in the present paper we are interested in the models that explore located latent classes within the Rasch family of models (named after Georg Rasch, 1960/1980, 1961), wherein there exists a sufficient statistic (generally the total score) for the latent variable. The existence of this sufficient statistic means that the models can be written in a conditional form, removing the person parameters (by conditioning on the sufficient statistic), allowing the estimation of item parameters without distributional assumptions on the persons (Rasch, 1961, 1968). Describing an analysis using the partial credit model (originally presented in Masters, 1982), Masters (1985) writes that "in practice, this latent trait analysis defines a limited number of ordered latent 'classes' and scales these classes" (p. 76). Similarly, Andersen (1988) discusses what he refers to as a "pseudo-latent-class model" that is the conditional form of the Rasch model. Clogg (1988) asserts that the conditional form of a Rasch model is a latent class model with one class for each total score. He calls this discrete version of Rasch model a discrete latent trait model. Later work refers to this model as the latent class Rasch model (Formann, 1995; Formann & Kohlmann, 1998).

Through an empirical example, Clogg (1988) found that "the scaled-latent-class model turns out to be equivalent to Rasch's model when the number T of latent classes is greater than or equal to $(J+1)/2$, where J is the number of items (p. 201)". This was formalized by Lindsay, Clogg, and Grego (1991; see also independent work by Follmann, 1988, and De Leeuw and Verhelst, 1986) who found that under regularity conditions, the latent class model with the class-specific response probabilities restricted as in the Rasch model is equivalent to the conditional Rasch model when the number of classes $T \geq \frac{J+1}{2}$. This is because for T large enough, the restricted latent class model exactly fits the observed score groups and the likelihood stops increasing. While model fit and statistical inference are the same for the Rasch model (estimated by conditional maximum likelihood) and the latent class Rasch model with enough classes, it is important to note that the mathematical form and conceptual framework remain different (Clogg, 1995).

There have also been a number of attempts to find a more general, common statistical framework for latent class and latent trait models. The main work around

these attempts can be broadly classified into two approaches: (1) log-linear models (70s-80s), and (2) generalized latent variable models (90s-00s). The goal under both approaches was to find a most general model for the analysis of latent variables.

The earlier common framework situated latent trait and latent class models within a log-linear formulation. In their most basic form the log-linear approach models the natural logarithm of the expected cell frequency in a contingency table as a function of a linear combination of response levels that define the contingency table (Agresti, 2007). Haberman (1979) formulated latent class models as log-linear models; Tjur (1982) and Kelderman (1984) showed how the Rasch model could be parameterized as a log-linear model (relying on the property of sufficiency).

This framework has been used to make comparisons between latent trait and latent class models (e.g. Clogg, 1988). Heinen (1996) uses the log-linear form to demonstrate equivalence between certain latent trait models and restricted latent class models, in particular when the estimation method for the latent trait model discretizes the latent variable. This parallels the preceding discussion with the common model framework abetting the comparison. Thus did Langeheine and Rost (1988), in their introduction to their volume on latent structure analysis, claim that "the log-linear model structure with latent variables is the joint supermodel of latent trait and latent class analysis" (p. 3).

More recent efforts to unite these models under a statistical framework can be grouped under the general conception of generalized latent variable models. These frameworks include the Generalized Latent Trait Models (Moustaki & Knott, 2000), Generalized Linear Latent Variable Models (GLLVM; Bartholomew & Knott, 1999), a generalized structural equation modeling approach (generalized SEM; Muthen, 2002) and the Generalized Linear Latent and Mixed Models (GLLAMM; Skrondal & Rabe-Hesketh, 2004) that encompass latent structure analysis as well as other models.

These frameworks focus on the generality of the models and usually consider a wider range of models. They aim to link specific models more broadly to the statistical literature and literature from other disciplines. For example, the frameworks draw on generalized linear (and non-linear) models and structural equation models. In doing so, the frameworks illustrate the generality and flexibility of the latent variable modeling approach. They encompass manifest and latent variables that are nominal, ordinal, and quantitative and include a variety of distributional links, mainly from the

exponential family. They also emphasize model extensions and the ability to realistically model complex field data that involve latent variables. These frameworks make clear that a wide range of latent structure models are related and can be compared.

In this paper, we focus on the links across latent structure models and comparisons among them. We therefore rely on the previous work that links the various models. We present and discuss the models using a common notation that reflects the common statistical frameworks within the generalized latent variable modeling realm.

3.2.3 Review of latent structure models

This section presents an overview of models within the framework of latent structure models. Common notation is used for all models. We consider P persons, indexed by p, who respond to I dichotomous items, indexed by i. The manifest item response of person p to item i is given by x_{ip}, where $x_{ip} = 1$ if the correct response is given and $x_{ip} = 0$ otherwise. For simplicity we present the models for dichotomous items; for most models, the extension to polytomous items is straightforward[2]. As mentioned in Section 3.2.1, all models share the assumption of local independence and our overview will focus on unidimensional models.

Classifying: Latent Class Models

Latent class analysis is concerned with grouping, or classifying, persons. The most basic form of latent class analysis addresses the first of the three kinds of latent structure, where the inter-individual differences are modeled as qualitative differences. The different latent classes in the model represent different 'types' of respondents as represented by panel (e) of Figure 3.1.

Latent class models assume that the differences between respondents can be modeled as a function of a mutually exclusive and exhaustive set of respondent 'types' (Heinen, 1993; McCutcheon, 1987). All the respondents of the same type (i.e. class) are assumed to be homogeneous in that they share the same probabilities of responding correctly to the items (i.e. they have the same expected value for each item). Accordingly, for this kinds of models the local independence is conditional on the

[2]However, as Croon (2002) points out, the polytomous extension of ordinal latent class models is not trivial.

classes, which means that within a latent class the manifest variables (i.e. the item responses) are statistically independent (Heinen, 1993).

The formal definition of a latent class analysis, which considers a categorical latent variable with C different groups, or classes, indexed by c, relies on two kinds of parameters to model the observed data:

$\pi_{i|c}$ The conditional probability of a person in class c correctly answering item i, formally defined as:

$$\Pr(x_{ic} = 1|c) = \pi_{i|c} \tag{3.1}$$

π_c The probability that a persons belongs to class c in the population.

Assuming local independence, the model defines the probability of a given response pattern for person p as:

$$\Pr(\mathbf{x}_p) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} \pi_{i|c} \tag{3.2}$$

In terms of the assumptions of the latent class model, it is worth highlighting that this model does not assume any kind of monotonicity (neither for person nor items), and relies solely on the assumption of local or conditional independence.

In the basic, unconstrained latent class model (Equation 3.2), the $\pi_{i|c}$ are free parameters, allowing a class-specific response probability for each item. The class proportions π_c, on the other hand, require a constraint, usually such that $\sum_c \pi_c = 1$.

We are particularly interested in the class-specific item response probabilities $\pi_{i|c}$. These can be parameterized using a logistic transformation:

$$\text{logit}[\Pr(x_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] = \beta_{ic} \tag{3.3}$$

This parameterization ensures that the probabilities $\pi_{i|c}$ fall on the interval $[0, 1]$ while allowing the estimated parameters β_{ic} to fall on $(-\infty, \infty)$. This parametrization is particularly useful for the comparison of latent class models to ordered latent class and latent trait models, where the parameter also typically falls in this range.

Additionally, this re-parameterization allows the β_{ic} to be defined as a linear combination of fixed, a priori effects which would correspond to the linear logistic

formulation for latent class models for both dichotomous (Formann, 1982) and polytomous items (Formann, 1992). This formulation also allows a number of extensions to the basic unconstrained latent class model such as equal class probabilities or equal item response probabilities (Formann, 1985, 1989).

Although the linear logistic formulation can support many potential extensions to latent class models, making them very flexible, in the context of this paper we are interested in the simple unconstrained latent class model. The reason for focusing on this particular model is the fact that there are no restrictions on the class-specific item response probabilities or on the class membership probabilities. Therefore, no ordering of either classes or items is considered, addressing the cases in which the structure of the latent variable of interest defines differences of kinds.

Ordering: Non-Parametric IRT and Ordered Latent Classes

The next group of models that we consider focuses on the models for variables that have an ordinal structure. In general, the models described in this section center around the idea that it is possible to rank both persons and items along an increasing progression of proficiency levels, where the positions on this progression are entirely defined by inequalities. Moreover, some of these models allow for the simultaneous ranking of both persons and items, indicating that the two rankings are consistent under the same proficiency progression.

Models that explore latent variables with an ordinal structure have been pursued in two different, albeit related, research traditions: non-parametric item response theory (NIRT; Sijtsma, 1998; Stout, 1987, 2001) and ordered latent class analysis (OLCA; Croon, 1990). Although OLCA was proposed outside the framework of NIRT, the connections between these two traditions have been explored in the literature where, for example, ordered latent class models have been reformulated as a special case of NIRT (Hojtink & Molenaar, 1997).

Non-parametric item response theory (NIRT) can be traced back to the work of Mokken (Mokken, 1971; Mokken & Lewis, 1982). Mokken originated the work in this field with his descriptions of two models that serve as the basis for NIRT: the Monotone Homogeneity Model (MHM) and the Double Monotonicity Model (DMM). MHM is concerned with the ordering of persons while the DMM is concerned with the simultaneous ordering of both persons and items.

The additional structure provided by models in NIRT comes at the cost of making more assumptions than for latent class analysis. According to Sijtsma and Molenaar (2002), the assumptions behind NIRT are:

(i) Unidimensionality, which means that the relation between the manifest variables depends only on a single latent variable.

(ii) Local independence, which in this case is conditional on the proficiency of the respondent θ, and is formalized in the models as:

$$\Pr(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{I} \Pr(X_i = x_i|\theta) \tag{3.4}$$

(iii) Monotonicity of the item response functions (IRFs): the ordering of persons is formalized by restrictions on the IRFs using inequality constraints such that:

$$\begin{aligned} &\text{if } \theta_a \leq \theta_b, \text{ then:} \\ &\qquad \Pr(X_i = 1|\theta_a) \leq \Pr(X_i = 1|\theta_b) \end{aligned} \tag{3.5}$$

These three assumptions are enough for applications of NIRT where the primary concern is the order of the persons, as in the case of the MHM. However, if the order of items is also a concern, as in the DMM, a fourth assumptions must be added (Sijtsma & Molenaar, 2002):

(iv) Non-Intersecting IRFs, which corresponds to the inclusion of item invariant ordering in the model. If $\Pr_i(\theta)$ is the probability of correctly answering item i by a person with proficiency θ, this assumption can be formalized as:

$$\Pr_1(\theta) \leq \Pr_2(\theta) \leq \ldots < \Pr_k(\theta) \text{ for every } \theta. \tag{3.6}$$

The key ideas from NIRT in the context of this paper are the emphasis on monotonicity and double monotonicity in the models. The MHM, through its assumption of monotonicity, provides the basis for ordering the respondents by increasing proficiency. The DMM, by the addition of invariant item ordering to the monotonicity assumption, allows applications such as the evaluation of theories about the cognitive complexity of the tasks and the identification of differential item functioning (Sijtsma

& Molenaar, 2002). These properties allow the exploration of a variety of research questions as well as application to practical contexts in which differences in ranking suffice to make judgments. This illustrates that the focus on models with an ordinal structure is a valuable practice in its own right, not simply as a stepping stone for models that assume quantitative structure.

Ordered latent class analysis (OLCA) is also concerned with the analysis of latent variables with ordinal structures. Within this tradition, it is possible to maintain focus on the properties of monotonicity and invariant item ordering while also taking advantage of a model-based analysis.

OLCA shares with LCA the assumption that respondents belonging to the same class have the same probabilities of responding correctly to the items, and that the 'types' considered in the model are sufficient to account for all the differences between the respondents. The formal definition of OLCA is the same as the definitions presented in Equations 3.1 and 3.2, where a categorical latent variable with C different groups, or classes, indexed by c, accounts for the relation of manifest responses as a function of two kinds of parameters, the proportion of persons in each class, π_c, and the logit transformation of the class specific probability of answering item i correctly, β_{ic}.

The difference between OLCA and LCA lies in the use of inequality restrictions that impose an ordinal structure on the relations between the β_{ic} parameters in the models (Croon, 1990, 1991). This is done "in such a way that one may say that the probability of a positive response increases when one runs through the set of latent classes from the 'lowest' to the 'highest' one" (Croon, 2002, p.140).

A model with ordered latent classes can be formally defined as:

$$\begin{aligned} \text{logit}[\Pr(x_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] = \beta_{ic}, \\ \beta_{ic} \leq \beta_{ic'} \text{ for all } c < c' \text{ and for all } i \end{aligned} \tag{3.7}$$

Additionally, OLCA offers the possibility of exploring the existence of invariant item ordering without class monotonicity by imposing inequality constraints according to the items instead of the classes:

$$\begin{aligned} \text{logit}[\Pr(x_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] = \beta_{ic}, \\ \beta_{ic} \leq \beta_{i'c} \text{ for all } i < i' \text{ and for all } c \end{aligned} \tag{3.8}$$

Finally, it is also possible to specify a model with double monotonicity by simultaneously applying the inequality constraints to both classes and items.

$$\begin{aligned} \text{logit}[\Pr(X_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] &= \beta_{ic}, \\ \beta_{ic} \leq \beta_{ic'} \text{ for all } c < c' &\text{ and for all } i, \\ \beta_{ic} \leq \beta_{i'c} \text{ for all } i < i' &\text{ and for all } c \end{aligned} \tag{3.9}$$

Based on these definitions it is possible to see that an ordered latent class model as defined in Equation 3.7 assumes[3]:

(i) Local Independence
(ii) Monotonicity in the classes

That the OLCA model defined in Equation 3.8 assumes:

(i) Local Independence
(ii) Monotonicity in the items

And finally, that the OLCA model defined in Equation 3.9 assumes:

(i) Unidimensionality
(ii) Local Independence
(iii) Monotonicity in the classes
(iv) Monotonicity in the items

Despite the considerable differences between the structure that is being specified by these three models, the models all have the same number parameters; moreover, they share the same number of parameters as the unconstrained latent class model. This is relevant when, for instances, one attempts to compare or select between these models using information criteria.

One issue that differentiates the OLCA approach presented in this paper from the traditional methods within NIRT is that OLCA is a model-based approach whereas NIRT

[3] A similar analysis of the assumptions of the MHM, DMM, and the Rasch Model from the perspective of NIRT is presented by Meijer, Sijtsma, and Smid (1990); the main difference with the one presented here is that in NIRT a θ is assumed in MHM, which is not the case in OLCA. This leads to Meijer et al. (1990) including unidimensionality as an assumption of the MHM, which is not included here for the person-monotonic ordered latent class model.

has a descriptive focus for determining model fit. Practitioners of NIRT usually rely on the calculation of scalability coefficients such as Loevinger's H (Loevinger, 1947, 1948) to asses the fit of the MHM model or the P(+ +)/P(- -) methodology (Sijtsma & Molenaar, 2002) for the exploration of non-intersecting IRFs in the case of the DMM model. In contrast, using OCLA, we can obtain a probabilistic, model-based evaluation of the properties of monotonicity, invariant item ordering and double monotonicity. As Van Onna (2004) points out, OLCA models have two advantages relative to NIRT in that they (i) allow the parameterization of the IRFs while remaining restricted only by inequalities and (ii) allow for reliance on statistical methods in order to check the adequacy of the models.

Another important difference between NIRT and OLCA, and one that is particularly relevant in the context of this paper, is the assumed latent structure behind the ordinal structure formalized in the models. Although the two NIRT models that have been described operate on the basis of inequality restrictions and will produce purely ordinal results, the models in fact assume that the underlying latent variable, θ, is continuous. This raises an inconsistency at a conceptual level between the NIRT approach to models for ordering respondents and the framework developed in this paper. According to the framework that is being presented, our interest is in adequately reflecting the structure of the latent variable. If we are interested in modeling the structure of the variable as ordinal either because our theory indicates so or because those are the kinds of inferences we intend to make, what is the advantage of modeling it assuming that the variable is continuous? Although it is possible to consider OLCA as special case of NIRT (Croon, 2002; Hojtink & Molenaar, 1997), the two approaches represent different conceptions of the latent structure.

It is important to highlight this point because ordered latent class models offer the possibility of conducting assessments that are consistent between the kind of predictions that most theories in the social sciences can afford and the formal model used to test them. In fields like psychology, sociology, or political science, the theories that are usually explored using quantitative models are more often than not incapable of generating truly quantitative predictions.

A fairly straight-forward counterargument to this view is that the use of quantitative models is acceptable because, although we are unable to make truly quantitative statements now, the quantitative structure of the variable will eventually be elucidated.

However, as Michell (1990, 2008b) has emphasized, this misses the critical point of questioning whether the variable that we are interested in assessing is quantitative at all (from his realist standpoint), or from the perspective of the PPM, considering whether we can justify such a model of the relevant attribute and whether the inferences we want to make regarding such structure can be justified for a specific context of application.

In many cases in the social sciences, neither the theories nor the predictions are expressed formally at a quantitative level and the assessment instruments are not usually tested to see if they justify quantitative inference. Faced with this lack of support, the prevalence of methodological models that assume a quantitative structure is surprising, pointing to either a a very strong conviction (or alternatively a highly unconscious or automatic one) that a quantitative structure is the best model for any relevant attribute.

On the issue of criticisms on the quantitative tradition in psychology, it is worth highlighting however that the perspective presented in this paper is in general consistent with the criticisms made by Michell (1999, 2008a, 2008b) that:

- psychologist (and arguably other quantitatively oriented social scientists) do not actively examine their assumptions regarding the quantitative structure of the attributes they study.
- there is a dearth of quantitative theories in psychology.
- quantitative social science tends to use numbers as a way of legitimizing its research results, which despite being often presented with decimal places, are not interpretable anywhere near that level of precision.

However, although we agree with Michell regarding these issues in the use of measurement instruments, the perspective presented in this paper and in the previous one present stark contrasts regarding both the source of the criticisms and the possible solutions. With respect to the source or rationale for the criticisms, Michell bases his criticisms from the perspective of metaphysical realism and it's concern for ontology, as opposed to the Pragmatic Perspective advocated in this and the previous paper, where "the practical bearings"—in the words of Peirce (1878/2014)—are the chief concern. In relation to the possible solutions to these issues, Michell advocates

that as the only alternative the thorough embrace of the Classical Definition of Measurement (Michell, 1999), metaphysical realism (Michell, 2005), and the methods of the representational theory of measurement (Michell, 1990), as opposed to the more encompassing definition presented in the PPM and the use of latent variable models as illustrated in the model selection framework discussed here.

Moreover, although Michell and others (Schönemann, 1994; Trendler, 2009) have focused their criticisms in the literature regarding the unquestioned assumption of quantitative structure of the relevant attributes, it is in principle equally problematic to blindly adopt any other assumption regarding the structure of the relevant attribute. After all, it seems similarly unwise to ignore evidence that an attribute behaves quantitatively as it is to ignore evidence that it behaves as a qualitatively distinct one.

Indeed, it may turn out that many of the attributes that are being studied are being appropriately modeled in terms of classes or in terms of quantities by social researchers.. Notwithstanding this possibility, it hardly seems rational to learn if this is the case by neglecting to conduct an active and explicit examination on the issue (and therefore not collecting evidence to justify the selection of one structure over others), and by this omissionignoring the fact that the quantitative, ordinal or classificatory structure of these variables has not been established. It is of the utmost importance to remember that these issues are not associated solely with the unexamined assumption of quantity, but in general with the adoption of any latent structure, for instance a classificatory structure, without an active examination of the matter.

In this paper, we rely on ordered latent class models for the assessment of ordinal differences. We purposefully wish to take a model-based approach. More importantly, we consider the fact that a continuous latent variable is not a necessary aspect of the conceptualization of OLCA a distinctive and valuable feature of these models. With the formulation based on latent classes, it is perfectly possible to interpret the results within a theory that simply specifies differences in ranking. Conceptualizing the variable in this way can better reflect the ordinal structure of a theory.

Quantifying: (Some) Continuous Latent Variable Models

The final set of models that we will review are models that allow the measurement of quantitative differences between persons. In order to clarify our understanding of

models that characterize the latent structure as quantitative, it is necessary to define two elements: (i) under what circumstances we consider that a variable is quantitative and (ii) what properties a model must posses in order to support the measurement of such a variable.

To address the first issue, we start from the properties that define ordinal structure. As Michell (1990) indicates, although all variables that are quantitative satisfy the conditions of simple order, not all simply ordered variables are quantitative. Beyond simple order, quantity involves additivity, which is a relation involving three elements in the form 'A + B = C' that must satisfy a specific set of conditions under the theory of fundamental and derived measurement (Campbell, 1928).

Although the definitions offered by the theory of fundamental and derived measurement has served measurement well in contexts such as physics, both theories ultimately relied on the existence of empirical concatenation operations as the source of the aforementioned property of additivity required to establish quantitative relations between objects (e.g. our ability to take two pieces of rope, put them end to end and observe that their combined length is indeed equal to the sum of each one of them) (van der Linden, 1994). The lack of the empirical concatenation operations in the social sciences has made the identification of attributes that directly satisfy these properties less than successful (Michell, 1990).

In order to address the need for quantification in the social sciences, Luce and Tukey (1964) developed the theory of conjoint measurement, which allows the determination of quantitative structure based on ordinal relations without the need for concatenation operations (Krantz et al., 1971/2007). This theory clearly defines what is a quantitative structure and specifies the conditions that we can test in order to confirm it.

Generally speaking, the theory of conjoint measurement establishes a set of conditions that if satisfied imply that three variables involved in a conjoint system of the forms presented Equation 3.10 it means all of them have quantitative structure (Michell, 1990; van der Linden, 1994). In the case of the Rasch model, the three variables involved would correspond to person proficiency, item difficulty and the odds of a correct response, and we are interested in determining if these three variables have a quantitative structure, and as Michell (1990) indicates, the theory of conjoint measurement is directly applicable to the situations in which none of the three variables are already quantified.

$$A = B + C \text{ or } A = B \times C \tag{3.10}$$

However, there is still a problem with the definition presented by conjoint measurement in the context of the varied assessment conditions and instruments used in the social sciences: the definition is deterministic. The use of a deterministic definition as opposed to a probabilistic one implies that any dataset that presents a violation of the conditions would fail to be considered as quantitative. This seems to be an unnecessarily stringent standard for datasets generated with instruments that are usually considered to capture not only the variable of interest but also measurement error.

A potential solution for this problem is to interpret the conditions and relations specified in the theory of conjoint measurement as referring to the model parameters in probabilistic models, which would afford the use of the definitions provided by that theory while at the same time allowing for the use of probabilistic models (Borsboom, 2005; Perline et al., 1979). If this interpretation is adopted, then models such as the Rasch Model (Rasch, 1960/1980) or the ADISOP models (Scheiblechner, 1999) would be correspond to what has been called a 'probabilistic variant' of the conditions imposed by the theory of conjoint measurement (Borsboom & Mellenbergh, 2004).

In this paper we will focus in the Rasch model to measure differences of quantity when the structure of the variable is considered to be continuous. According to Meijer et al. (1990) the Rasch Model relies on five assumptions:

(i) Unidimensionality
(ii) Local Independence
(iii) Monotonicity in θ
(iv) Monotonicity in δ
(v) Minimal Sufficiency

As Meijer et al. (1990) indicates "[The Rasch Model] is based on the same set of assumptions as the Mokken model of monotone homogeneity, plus the assumption of minimal sufficiency of the unweighted person and item sum scores for the estimation of the θ and δ parameters, respectively" (p. 284). With these assumptions, the Rasch model attempts to model the correlation among the manifest variables as a function of the difficulty of the items and the proficiency of the respondents. The ability of

separating item parameters δ_i from person parameters θ_p is one of the most useful aspects of the Rasch model and lies at the center of Rasch's definition of specific objectivity (Rasch, 1961, 1968). This is in contrast to the joint parameters, β_{ci}, that were used in all of the previous models.

Formally, the Rasch model can be written as:

$$\text{logit}[\Pr(x_{ip} = 1|\theta_p, \delta_i)] = \theta_p - \delta_i \tag{3.11}$$

According to Rasch (1968) the separability of parameters is the key feature that, in principle, allows measurement in the social sciences and humanities in the same sense as in physics.

Specific objectivity, and parameter separability by extension, are important as measurement properties because they imply that the differences between any two persons or objects are independent of the questions or objects that were used to compare them and vice versa (Rasch, 1968). A violation of this property would imply, for example, that the length of objects A and B would seem to be $A > B$ under one measuring tape and $B > A$ under another; a results that clearly violates our intuitions about what constitute a valid measure. Adherence to this property, implies that only models within the Rasch family can serve as the basis for measuring variables with quantitative structure, thereby excluding models with unequal factor loadings such as the two parameter and three parameter logistic family of models.

It is important to note that is not the only response to the challenges raised by models that allow unequal factor loadings; as Borsboom (2005) argues, it is also possible to extend our definition of measurement in order to include the possibility of violations like the one just described. Borsboom's point is that although the previous example is clear and intuitive in the context of what we know about physical sciences, this need not be the case for social sciences. After all, "...when we move from the physical to the psychological world there is no reason to suppose that the methodological requirement continues to hold, because substantive considerations need to support it" (Borsboom, 2005, p.117). However, it is our position that changing the idea of measurement by accepting violations of specific objectivity is not a mere extension of the meaning of measurement but a significant, and potentially misleading, redefinition of the concept. If we allow such a redefinition, it would mean that when

someone says 'I measured A and B and I know that A is greater than B' we would need to ask, did you measure with an additive or a non-additive model? in order to make sure to what extent A is greater than B. This does not mean that non-additive models are in principle wrong or that they do not have a place in science, only that the use of non-additive models produce results that are methodologically so dissimilar that it merits the use of a different word than "measurement"[4].

In this paper, we are adopting a formulation of quantitative structure based on the theory of conjoint measurement, but we are interested in exploring its extension as probabilistic models. Although Rasch based his definition of measurement on the idea of *specific objectivity*, the definition offered by the theory of conjoint measurement is consistent with the overall measurement perspective developed by Rasch (1961, 1960/1980). The formal similarities between the Rasch model and the definitions of conjoint measurement have led a considerable number of researchers to state that the Rasch model is conjoint measurement or a probabilistic version of it (Borsboom & Mellenbergh, 2004; Embretson & Reise, 2000; Kline, 1998; Perline et al., 1979; Scheiblechner, 1999). However, whether this interpretation is valid has been questioned from several perspectives (Kyngdon, 2008; Michell, 2008a).

For the purpose of the present paper, the main issue with the interpretation of Rasch modeling as conjoint measurement is that although conjoint measurement specifies conditions that can be tested in order to assess quantitative structure, the so-called probabilistic version of it, the Rasch model, actually assumes a quantitative structure. In other words, the use of the Rasch model does not offer any assurance that the structure of a variable is quantitative, because it is assuming what we want to determine from the start. This issue is clear in the common methods in Rasch measurement (and IRT) of testing items for detecting misfit. These methods, using such indicators as the mean square statistic, are based on a conception of the variable that is quantitative, and hence cannot, in their current form, test for this assumption.

However, the problem is that unless we contrast the Rasch Model with other models that offer alternative latent structures, it is not possible to test the fundamental assumption about the quantitative structure of the data. While there are various types

[4]If Borsboom is right and the objects of study in the social sciences are so radically different from the ones studied in the physical sciences so that their study merits the abandonment of the constraints that we commonly attribute to the concept of measurement, it would seem reasonable to describe them using a different methodological framework.

of diagnostic misfit indices available for the Rasch Model, these do not directly test the assumption of continuous latent structure, but only assess the adequacy of the functional form specified by the model.

In order to address this problem, the framework proposed in this paper will rely on a comparison between the ordinal double monotonicity model (Equation 3.9) and the Rasch model. However, a direct comparisons between the double monotonicity model and the Rasch model would not only test for the difference of scale, but also would involve the issue of the appropriate determination of the number of classes. It is possible to solve this issue by relying on the latent class Rasch model (Formann, 1995; Formann & Kohlmann, 1998). As discussed in Section 3.2.2, the fit of a latent class Rasch model with the number of classes equal to at least half the number of items is equivalent to the fit of a Rasch model (Lindsay et al., 1991). Comparing the latent class Rasch model to the double monotonicity latent class model isolates the assumption of a continuous scale as the only difference between the two models.

The key difference between the Rasch model and the latent class Rasch model corresponds to the number of proficiency parameters that are estimated. Where the Rasch model allows, conceptually, one proficiency for each student, the latent class Rasch model assumes that the respondents are grouped in C classes and that all the members of a class share exactly the same proficiency. Of course, we can specify the latent class Rasch model to have potentially as many classes as total scores in an instrument, making it effectively equivalent to a Rasch model. This equivalence happens because the Rasch model only allows one proficiency for each sum score (assuming no missing data), so that the effective subscript for the ability parameters is by score group. This functional discreteness is the basis for the equivalence between the models discussed in Section 3.2.2. In general, however, the number of assumed classes for the latent class Rasch model is less than the number of possible total scores, so that the latent class Rasch model conceptually encompasses a discrete latent variable while the Rasch model does not.

The formal expression of the latent class Rasch model makes clear that the only difference between the formula for the conditional probability of a correct response in the two models resides in the subscript of the proficiency parameter θ, which corresponds to a person p in the case of the Rasch model (or, in practice, to a total score t) and to a class c in the latent class Rasch model. The latent class Rasch model is formalized as:

$$\text{logit}[\Pr(x_{ic} = 1|\theta_c, \delta_i)] = \theta_c - \delta_i \tag{3.12}$$

Additionally, the latent class Rasch model shares with the other latent class models the need for a set of C parameters π_c that represent the estimates of the proportion of respondents that belong to each class in the population.

3.3 Model Comparison Framework for Selection of the Latent Structure

3.3.1 Models and their Assumptions

According to the measurement definition presented in the PPM, measurement is an activity of classification, ordering or quantification. Under ideal circumstances we would expect alignment between our theoretical model, our statistical model, and our inferences to be well aligned, meaning that if the inferences are to be ordinal, we would have a similar theoretical and statistical model. However, the unexamined selection of a statistical model simply may lead to the formulation of inferences that are, from a statistical perspective, not supported by the available data. To put it simply, we could conceivably have a classificatory theory of a relevant attribute, say civic engagement, and we may want to make inferences regarding different classes of behavioral patterns to which respondents adhere in this attribute, but it might be the case that the data is simply inconsistent with this latent structure.

The lack of examination of the assumptions regarding the latent structure of variables in the social sciences has been forcefully criticized by Michell (2008b), Cliff (1992), and Schönemann (1994). The criticism of these authors has pointed towards the failure of social scientist to adopt the methods developed in the representational theory of measurement for testing the presence of a quantitative structure in the data, arguing that so far this is the only well established method for conducting such test in the social sciences (Michell, 2008b). However, it is important to remember that, from the perspective adopted in this paper, the adoption of a classificatory or ordinal structure needs to also be actively examined and supported by evidence.

In this paper we introduce a model selection framework that would directly address this concern, taking advantage of a variety of latent variable models to collect

evidence supporting the selection of a latent structure. In order to do so, we rely on the differences between quantitative, ordered, and qualitative latent structures, which can be represented by a series of the latent variable models discussed in the previous section that progressively assume a more constrained latent structure.

In order to capture the range of possible structures for the latent variable we will rely on six models:

(i) The Unconstrained Latent Class Model (UN).
(ii) The Ordered Latent Class Model with Class Monotonicty (MON).
(iii) The Ordered Latent Class Model with Invariant Item Ordering (IIO).
(iv) The Ordered Latent Class Model with Double Monotonicity (DM).
(v) The Located Latent Class Model or Latent Class Rasch Model (LCR).
(vi) The Rasch Model (RM).

These models are yet a smaller subset of possible mathematical models than the one represented in Figure 3.1, with the RM and the LCR as models that assume a quantitative structure, with the former associated with pane (a) of Figure 3.1, and the latter with pane (c) of Figure 3.1. Within the model selection framework presented here, there is no instance of models of located heterogeneous located classes corresponding to pane (b) of Figure 3.1. The models MON, IIO, and DM are three instances of the ordered latent class models presented in pane (d) of Figure 3.1. Finally, the UN model in this framework represents the types of models in pane (e) of 3.1. We can visualize the specific subset of models included in this framework in Figure 3.2.

The differences between these models can be accounted for by two assumptions, one of order and one of scale. For example, the MON, DM, LCR, and RM models all assume that the classes (or persons) are ordered, while the IIO, DM, LCR, and RM models all assume that the items are ordered. The MON, IIO, and DM models assume an ordinal scale, while the LCR and RM models assume an interval scale. The assumption of ordering can apply either to the persons/classes or to the items/tasks while the scale assumption (i.e. classificatory, ordinal, or interval) necessarily applies to both of them.

Figure 3.3 graphically depicts how these assumptions are reflected by the model parameters. Notice that the six plots present patterns that are illustrative of the log-odds of answering correctly 10 different questions under each of the different models.

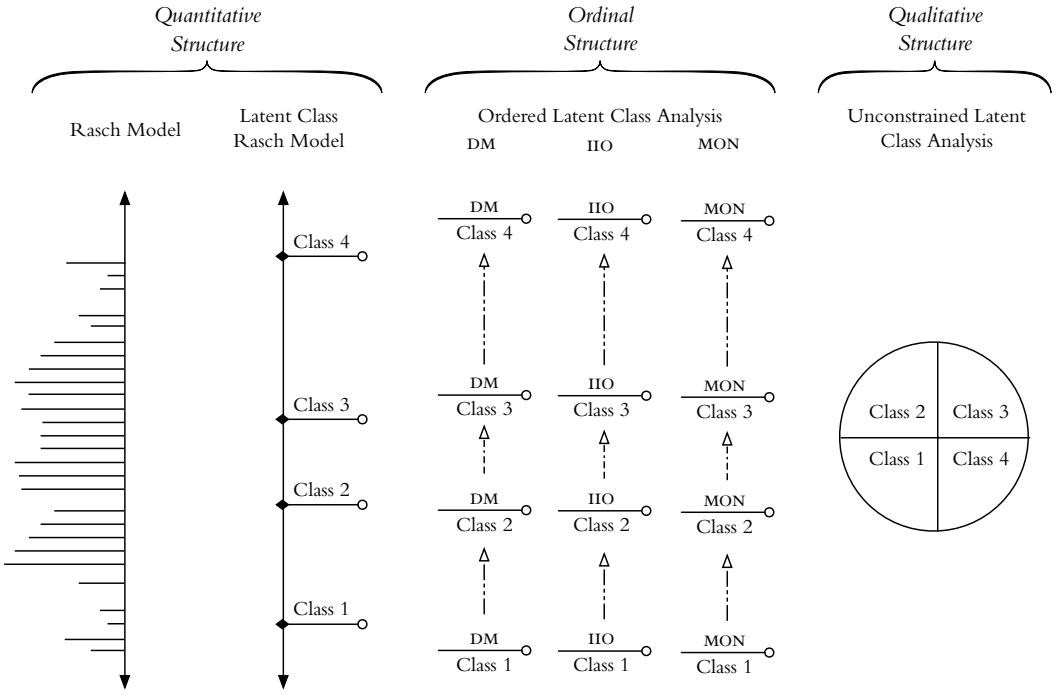


Figure 3.2: The Six Models Included as Part of the Model Selection Framework.

Despite their similarity, these plots are not item response functions, which are usually presented as a function of the person proficiency (θ), as not all the models include it as a parameter.

It is possible to see how the parameters in the unconstrained latent class model, represented by panel (a), do not show any recognizable structure. The class monotonicity model restriction in panel (b) are expressed in the fact that the lines that represent each class never intersect. In both these cases the x-axis formed by the items is purely qualitative and cannot be interpreted as the dimension of increasing proficiency.

In the case of the invariant item ordering model, shown in panel (c), the order restriction on the items is represented as a non-decreasing pattern in the lines of all the classes. The introduction of the invariant item ordering restriction in this panel allows the interpretation of the x-axis as a discretized representation of increasing proficiency, a characteristic shared by panels (c) to (f), making them very similar to a representation of a non-parametric item response function. The double monotonicity model restrictions represented in panel (d) build on the scale provided by the

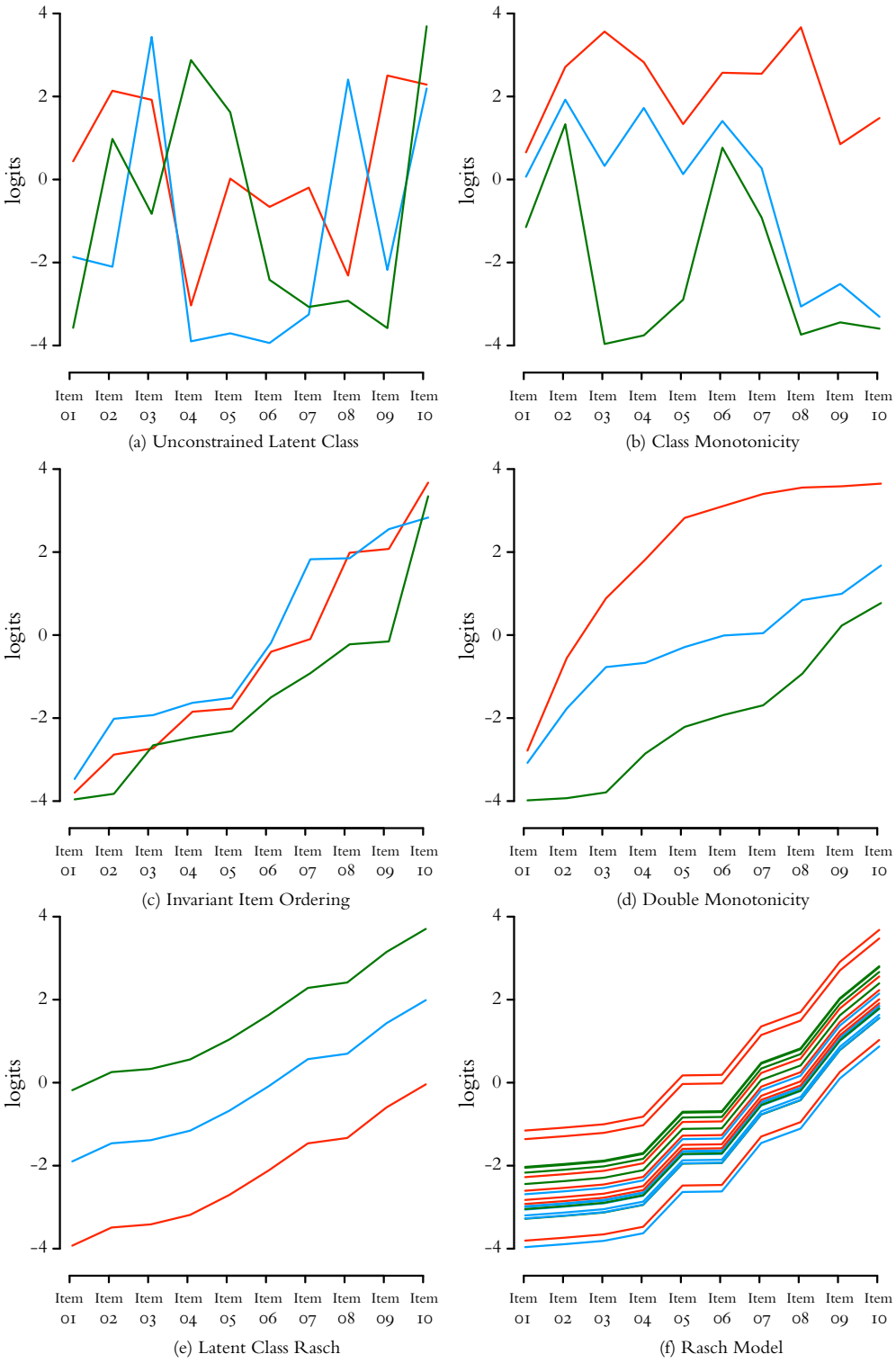


Figure 3.3: Relations Between the Six Models for Different Classes or Persons.

invariant item ordering restriction showing a non-decreasing and non-intersecting pattern of lines.

The last two panels, (e) and (f) show how the effect of the assumption of an interval scale in the Latent Class Rasch and the Rasch Model is expressed in the parameters. These two panels show the same properties as the double monotonicity model but they illustrate how the effect of each item in each class is systematic. In other words, it can be seen that the difference between the lines is constant, which corresponds to the differences in proficiencies between the classes in the case of the LCR and between the persons (i.e. different total sum score groups) in the Rasch Model.

The different assumptions regarding the latent structure determine the different kinds of constraints placed on the model parameters, which in turn formalize the different structures for the latent variable. The comparison between these six models will allow us to "decompose" the scale and order assumptions in order to evaluate the extent to which is valid, from a statistical perspective, to adopt these assumptions. As a direct result, considering these models can help us make a more informed decision regarding the statistical model (and its implied latent structure), and their comparison can be used as a source of evidence to support their selection.

Table 3.1 presents a summary of the assumptions of each model and the kinds of inferences supported by each model.

Table 3.1: Summary of Assumptions and Inferences Associated With Each Model

	UN	MON	IIO	DM	LCR	RM
Assumptions						
Local Independence	Yes	Yes	Yes	Yes	Yes	Yes
Person Monotonicity	No	Yes	No	Yes	Yes	Yes
Item Monotonicity	No	No	Yes	Yes	Yes	Yes
Unidimensionality	No	No	No	Yes	Yes	Yes
Quantitative Scale	No	No	No	No	Yes	Yes
Inferences						
Distinctions of Kinds	Yes	No	No	No	No	No
Distinctions of Order	No	Yes	Yes	Yes	Yes	Yes
Distinctions of Quantity	No	No	No	No	Yes	Yes

As it is emphasized in the definition of measurement put forward as part of the PPM, the goals of measurement play a central role in the process:

> ...the goals provide the criteria for judging whether the measurement we are examining is a good one or a bad one depending on the extent that they help us accomplish them. Associating a measurement with a specific and explicit goal both (a) restricts its scope, hopefully tempering inferences and statements based on it, and (b) establishes the criteria by which the measurement is to be judged. (See Section 2.5.4)

If our goals demand from us ordinal inferences, it is natural to engage in measurement from the perspective of ordination as the activity that needs to be conducted, and hopefully, our model of the relevant attribute will be consistent with an ordinal interpretation of the results. However, throughout this paper we make the case that, when adopting a statistical/measurement model, it seems unreasonable to ignore the possibility that—although we want ordinal inferences and we have an ordinal theory—the structure of the collected data may not plausibly behave as data generated from an ordinal latent structure.

This framework is presented as a tool that researchers can use in order to empirically gather evidence of whether the collected data seems to be plausibly generated by the kind of structure that is both defined by the model of the attribute and the structure that more closely resembles the inferences that he or she wants to make. In this way, if the researcher initially would have adopted a DM model, this framework encourages he or she to compare it both with models that relax some of those assumptions (i.e. IIO, MON, and UN models), and with models that present further assumptions (i.e. LCR and RM); the former helping her or him determine the extent to which ordinal models are more plausible than unconstrained models and the latter informing her or him if a quantitative structure should also be considered.

This procedure opens of course the possibility that the empirical results do not align well with the initial goals and/or with the model of the attribute. A scenario such as this will surely present a challenge to the researcher, who will have to consider between competing hypotheses that can, for example, blame the inconsistency on the specific sample that was examined (e.g. "the sample was too homogeneous, it does not represent the variation in the population"), the instrument that was used (e.g.

"the questions were poorly formulated"), or potentially on the model of the attribute (e.g. "personality types seem to be unsuitable for ordinal classification"). In any case, finding this inconsistency should prompt the researcher to either redo one or more aspects of the measurement process, or potentially, to decide to move forward with the originally intended model knowing that the inferences that he or she will make may be inconsistent with the structure of the collected data, hopefully by making this explicit and putting forward a case to justify ignoring the empirical evidence on this point. These issues are discussed in more detail in the following section, which tackles the implications of this framework for theory development.

3.3.2 Implications for theory development

Up to this point, when discussing the different structures that can be defined for the differences among the persons or organizations being assessed, we have referred more or less explicitly to this structure in three different ways:

(a) The structure as defined in our theory about the relevant attribute.
(b) The structure formalized in our statistical model.
(c) The structure in the observed data.

An overly simplified and optimistic conception of the assessment process would indicate that if our theory of the structure of the relevant attribute is correct (understanding by this that our theoretical model allows us to make useful inferences in context of application defined by our goals), then the observed data should be a function of that structure and by formalizing that structure in our statistical model we could be able to test (i.e. attempt to falsify) our hypothesis about the data.

This understanding of the process is presented in Figure 3.4, where the structure of the relevant attribute is shown as directly determining the structure of the observed data, and where the statistical model is a formal expression of that structure. If this was the process of assessment, and we could then make a contrast between our expectations according to the model versus the observed data, represented by the double headed arrow that connects the formal versus observed structure, would allow us to directly test our theories about the structure of the data.

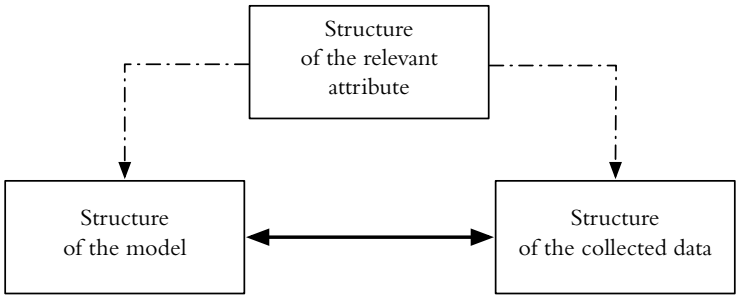


Figure 3.4: A Very Simplistic Representation of the Assessment Process

A slightly less simplified, and far less optimistic, conception of the assessment process also includes a fourth element as a potential source of structure in the data, namely:

(d) The structure as tacitly defined by the assessment instrument.

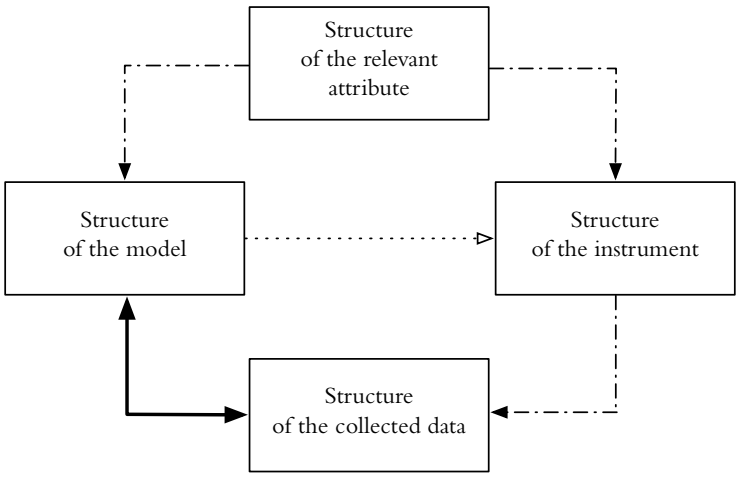


Figure 3.5: A Slightly Less Simplistic Representation of the Assessment Process

The introduction of the measurement instrument in Figure 3.5 highlights that, independently of what an idealized "true" structure of the relevant attribute, the good or bad fit between a model and the observed data can be attributed either to the presence (or lack of) structure in the instrument. Simply put, if we set up a statistical model for a quantitative latent structure but the model does not fit the data, we could

assume that our hypothesis about the structure of the relevant attribute is wrong, but it is also possible that the poor construction of the measurement instrument is the problem. In fact, in contrast to the approach advocated in this paper, in practice the measurement instrument is blamed more often than the hypothesized structure is questioned. We consider that it is important to actively recognize that both are competing explanations, and that the selection of one over the other should be made by considering both theory and evidence, and not by defaulting to one or the other.

In an ideal world, our assessment instruments would be 'transparent', producing data untainted by external influences. However, it seems plausible that our instruments can either obscure the structure of a variable or create an artifact by inducing structure that is not a characteristic of the variable itself. An interesting example of the latter is provided by Borsboom (2005) where he discusses how a set of questions that target inclusive subclasses of people can produce results that would match a Guttman scale, potentially leading to the inference of an ordinal structure out of qualitative differences; we can also consider the opposite case, where the structure of an instrument could induce a classificatory structure when a different instrument could have presented a quantitative one. On the other hand, it is always possible to argue that results that show consistency between the data and the statistical models are an artifact due to the specific tailoring of the instrument to a narrower scope of the attribute than what the theory implies—thereby tacitly changing the attribute's definition—in order to to fit the specifications of a statistical model. When we try to measure concepts that are complex, say "solidarity," "anxiety," "civic engagement," it is easy to ignore or remove certain aspects of the concept in order to make the instrument "behave."

Changing the definition of the relevant attribute in order to conform to the requirements of a given method is in principle fine, as it can be argued that in broad terms this is the way in which physics, for instance, "tamed" or "refined" the concept of temperature by stripping away initial intuitions about it, including its link to the human perception of warmth and coldness; in other words, the development of instrumentation interacts with the development of our theory of the attribute. However, for this to actually be a contribution, it is necessary to explicitly make the case that the new definition should overtake the previous, larger one, understanding that the instrument is a measure of this newer, more "refined," definition. Alternatively,

it would be necessary to acknowledge that the attribute A as originally or commonly defined is not suitable to be measured under a specific method, and therefore the instrument was modified in order to measure attribute A^*. This is a particularly complex issue in the social sciences, as researchers can make a series of tweaks to the measurement instrument without necessarily advocating the reconceptualization of the attribute based on those modifications. More importantly, these adjustments can significantly change the sense on which the attribute that is supposed to be measured is commonly used, without acknowledging that the scope of the attribute included in the instrument falls short of representing what is understood by the public. It is worth emphasizing however, that this process of revision of the construct A is not necessarily a negative process as long as any changes to the attribute are explicitly dealt with. When this is the case, this process can lead to productive and explicit revisions of the definitions of the attribute (or even to potentially make further distinctions such as decomposing parts of A into a distinct attribute B), as advocated in current instrument development frameworks such as Wilson's (2005) *Construct Modeling* approach and Mislevy's (2006) *Evidence Centered Design*.

Although it is true that this last representation is also a simplification of the assessment process, we believe that it is important to highlight the role that the assessment instruments can have both as a factor in the data generation process but also as an alternative source of explanation when interpreting results. This point can be considered a specific case of the much more general point of underdetermination of scientific theories as pointed out by the Duhem-Quine thesis (Quine, 1951) which indicates that a hypothesis can never be tested in isolation. In this particular case we are testing not only our hypothesis about the structure of the variable but also, at the very least, we are also testing our assumptions about the quality/validity of our instrument.

In light of the above points, it is important to clarify that the framework for tenable assessment proposed here can inform the development of theories about the structures of the variables by pointing out discrepancies, but it cannot be used to directly determine the validity of any particular theory about the structure of a relevant attribute. It is also worth highlighting that this complexities appear when we include just one element—the instrument—as an additional source of complexity; the PPM highlights additionally as fundamental the role of of the context of applications as a function of the overall goals that direct the process, which introduces additional

complexities that have been ignored so far by assuming that somehow we can hold the context constant.

3.3.3 The issue of grain size

The presence of the assessment instrument intervening between the latent variable and the data, and perhaps providing or masking structure, gives rise to a further complication if we just start to consider the role of the context. This complication is illustrated by means of the following example.

Consider the measurement of height. It is widely accepted that the structure of the "latent" variable is quantitative (and, in fact, is a oft-cited example of a ratio scale). In order to measure height, we must use an instrument, such as a ruler or tape measure. Measuring tools such as these are usually discrete; that is, they provide a measure of length according to some discrete unit such as inches or centimeters. If the "true" length of an object corresponds exactly to 64.5 centimeters it may be recorded as 64 cm or 65 cm. In addition, actual use of a measuring tool introduces measurement error. So, that same length could even be recorded as 62 cm or 68cm, depending on the quality of the measurement. Through this combination of measurement error and the discrete measurement tool, the recorded heights of a group of objects may not enable us to make valid quantitative inferences for each one of them.

However, we may still be able to get valid quantitative inferences between, say, groups of persons, for example between the two "latent" groups of people say, before and after puberty (with the former being shorter than the latter). Alternatively, we may be able to provide valid rankings of height even for individuals (so long as they are not too close together). This issue will occur even though we are measuring length, which usually serves as the definition of an interval scale!

Determining the scale of the latent variable is in this manner confounded by the grain size desired for the analysis, which in turn is defined by the goals of the measurement process. We may not be able to make tenable quantitative claims about length if the data is recorded in small units or if the measurement error is sufficiently larger than the unit of recording. However, even in these instances, we may be able to make tenable ordinal claims about the length of sufficiently different groups of objects.

This same issue can be illustrated by an example from Rost (1988, 1990). In these papers, Rost discusses the analysis of a 10 question physics knowledge test, which he determines is not Rasch scalable. In doing so, he also presents results from fitting latent class models with three and four classes to the data. In the three class model, a plot reveals that the item response functions do not cross, but in the four class model, the item response functions do cross. Thus, the three-class solution displays class monotonicity while the four-class solution displays no discernible ordered structure. So, if we wanted to consider the analysis of three classes, we could rank the latent groups of people (if we also determine that the three-class solution provides an acceptable fit to the data), but we could not order the latent classes if we decide on the four-class solution.

The issue of grain size as reflected by considering different numbers of latent classes is illustrated by Figure 3.6. In the figure, are there nine unordered groups of persons or three ordered groups of persons?

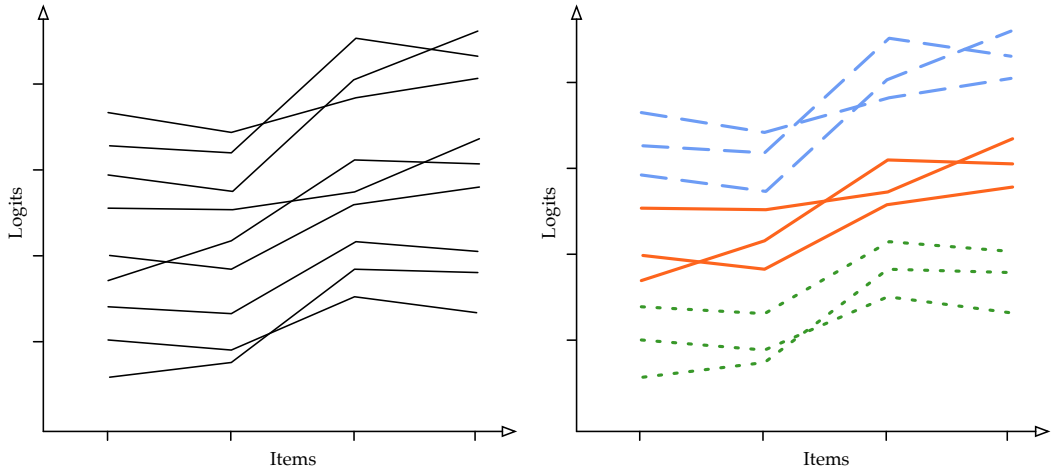


Figure 3.6: IRFs for Nine Unordered Groups or Three Ordered Groups?

The issue of grain size is also reflected in the work on invariant item ordering (IIO) by Ligtvoet (2010), who was primarily concerned not with the structure of the latent variable but with the structure of the items. He reported the results of a study regarding the sample sizes needed to infer that items are ordered such that IIO holds and concluded that for "reasonable sample sizes" (a few thousand) only up to 6 invariantly ordered items can be distinguished. In his discussion, Ligtvoet proposes

two alternative approaches to investigate IIO. The first is to consider clusters of items "with similar IRFs" and the invariant ordering of these item clusters. The second is to use a coarser division of the latent variable (i.e. use fewer latent classes). He suggests that both of these approaches might allow more powerful tests of IIO, though they may cause bias (distort the view of the IRF).

It is clear that the selection of a latent variable model and the use of model selection to test assumptions regarding the structure of the latent variable is complicated by considerations of the grain size of the claims we want to make based on our analysis of the latent variable. If the hypothesis of a particular latent structure is rejected, we can consider two options: (1) relaxing (or changing) the assumptions regarding the latent structure; or (2) considering a different number of groupings / classes. Though we wish in this work to focus on assumptions regarding the structure of the latent variable, it is important to note that we cannot do so without also considering the issue of the number of latent classes.

3.3.4 Model Selection

As indicated before, the general aim of this framework is to identify the most likely structure that generated the data, keeping in mind the previously discussed limitations for attributing any structure to the variable of interest. The identification of this structure is achieved by comparing the fit between the six aforementioned models and selecting the model that offers the best relative fit to the data. Using the model identified with this framework would allow us to make tenable claims about the kinds of inter-individual differences that can be detected in the data.

In a a typical research context it is natural to rely on existing knowledge (such as extant theories, prior data, or desired purpose) to guide the model selection process, for example by defining the set of models to be considered or by serving as evidence used to decide between models. The model selection process outlined here is purposefully freed from such a priori assumptions. Although previous knowledge undoubtedly plays an important role in the process of research, it can also be the source of serious blind spots. Therefore our focus here is to critically examine the assumed latent structure, suspending our reliance on our prior knowledge, in order to be open to the possibility that some of our assumptions may be incorrect. Even

in the case that the prior knowledge comes from a critical reflection on the latent structure, we still think it is beneficial to reexamine the latent structure in light of new data.

The strength of this approach relies on the successive comparison of models rather than a simultaneous comparison among the six of them. Although it can be informative to learn that an unconstrained latent class model fits comparatively better than the Rasch model in a dataset, the conceptual differences between the models shed little light regarding the patterns in the data that account for the fit differences. If we focus on the successive comparison between the models we will be able to learn specific aspects of the structure in the data.

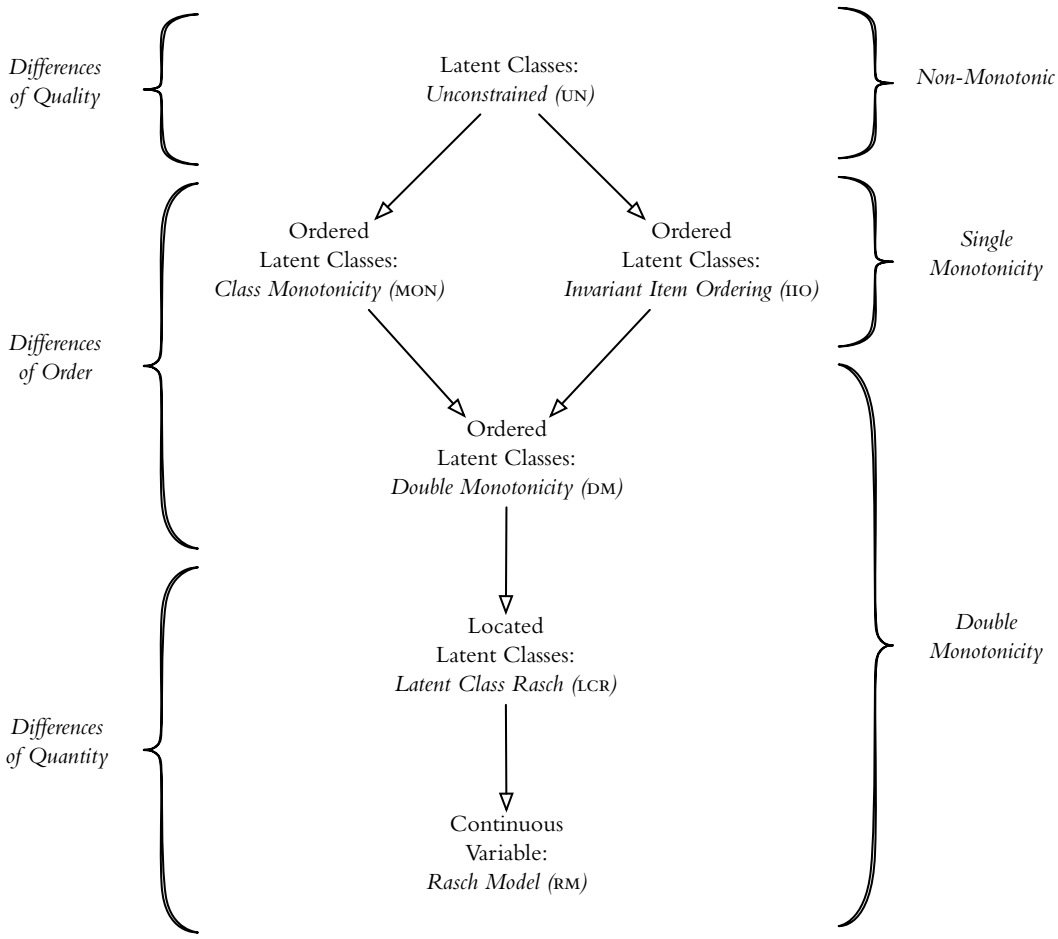


Figure 3.7: Diagram Depicting the Relations Between Models and Assumptions.

A summary of the six models that are being considered and how the monotonicity and scale assumptions are formalized in the models is presented in Figure 3.7, where the arrows indicate increasing structure.

The initial comparison between the unconstrained model and the ordered latent class models with single monotonicity informs us separately whether the data seems to support either of the two basic relations needed to think about a progression of levels of proficiency as opposed to somewhat incommensurable differences in "kinds" of performance, effectively differentiating between a classificatory and ordered structure.

The comparison between the single monotonicity models versus the double monotonicity model informs us if the differences in proficiency levels can be conceptualized as ordering the respondents into a single dimension of proficiency. This is because the proficiency classes detected in the class monotonicity model now share a similar relation with the proficiency progression contributed by the presence of invariant item orderings (i.e. in the sense that they are similarly correlated with the latent variable).

The next comparison of interest provides evidence of the presence of a quantitative structure in the data as opposed to an ordinal structure by contrasting the Double Monotonicity Model and the Latent Class Rasch model. The rationale for this is based on the fact that the only formal difference between these two models is the addition of parameter separability in the Latent Class Rasch model, from which follows the interval structure of the scale. This jump from ordinal to interval scale between models that assume double monotonicity has been explored for example by contrasting the DM Model to the Rasch Model, where the only difference in terms of assumptions is the presence of minimal sufficiency (Meijer et al., 1990), which is directly related to the separability of parameters that allows for the comparison of persons and items in an interval scale.

The final comparison between the latent class Rasch model and the Rasch Model provides information about the presence of a "grain size" issue as discussed in the previous section by indicating whether the Rasch model is to some extent overfitting the data by considering different locations in the proficiency continuum for each total sum-score detected in the dataset.

3.4 Discussion

The framework presented in this paper offers an alternative within the latent variable framework to contrast alternative latent structures in order to explicitly test the consistency between the data and the assumptions of an ordered or quantitative structure. The use of a single framework does not by itself definitely prove or disprove the validity of using a particular latent structure analysis to model data when measuring a relevant attribute; however, it does provide a source of evidence to support the adoption of one latent structure or the other, or alternatively, it provides diagnostic information that can inform and help revise the theory, the structure of the instrument or change the "grain size" of the inferences that are to be made.

Although the conceptual comparison between the models is fairly simple, the statistical comparison between these models (previously described as equivalent models in Section 3.2.2) is not as straightforward due to the fact that not all of them are nested in the traditional sense. More specifically, it is the case that four of the six models, namely the unconstrained model and the three ordered latent class models, will always share the same numbers of parameters. These models are nested in the sense that inequality constraints are applied to their parameters, but the number of parameters remains the same. This has implications for the use of information criteria for model selection because these four models are subject to the same penalties as a function of their total parameter count despite the fact that the double monotonicity model is applying important additional constraints to the structure of the parameters.

This issue, and certainly others, will need to be tackled in order to streamline and improve this framework, but the overall rationale for the comparison outlined in this paper offers a blue-print to directly addressing issues that so far are thought to only be examinable under the representational theory of measurement. The possibility of examining them under a latent variable framework can potentially, assuming the issues discussed in the previous paragraph are overcome, make this analysis commonplace, at least in practical terms, as most if not all of the models discussed in the framework can be estimated with general statistical software such as Stata, R, SAS, LatentGOLD and Mplus, with the estimation of IIO models being the exception.

The use of this framework to critically examine the assumptions regarding the latent structure that can be supported based on the data and, more generally, to

question and revise our assumptions regarding the structure that we ascribe to the relevant attributes that we study would be a significant step towards countering the "pathological behavior" criticized by Michell (2008b). Although it is unlikely that the method presented in this paper will resolve the objections raised by critics like Michell (2008b), Trendler (2009), or Schönemann (1994), its adoption as part of the psychometrician's armamentarium would address the accusation of actively ignoring the issue of testing the assumptions about the structure of the attributes under study.

Chapter 4

The Ordered Mixture Linear Logistic Test Model (OM-LLTM)

4.1 Introduction

Often we need assessments in educational contexts to go beyond how people performed and give us clues as to *why* they performed in that way. Although these two types of information are not necessarily incompatible, there is a tension between, providing easily interpretable results to stakeholders on the one hand, while on the other hand providing more fine grained diagnostic information.

A common way to communicate assessment results in a simple and understandable manner is the use of ordered performance levels, such that individual differences are characterized as membership in one of several ordered performance groups. The—sometimes brief—descriptions and labels of each of these levels play an important role in shaping inferences regarding student performance and capabilities.

When the focus is on the diagnostic use of an assessment, psychometric approaches have relied on the analysis of the content domains as a function of a set of basic features that are instantiated in characteristics of the tasks that are presented to the students (e.g. Diagnostic Classification Models; Maris, 1999; Rupp et al., 2010). Under this kind of model, each student is then associated with a specific profile in terms of these basic features.

This paper presents the Ordered Mixture Linear Logistic Test Model (OM-LLTM), a model well suited to simultaneously address these two demands, providing an over-

arching result expressed as the membership in an ordered class while at the same time providing information about the way students in each class perform on different item features.

The OM-LLTM combines the strengths of the linear logistic test model (LLTM; Fischer, 1973) and ordered latent class analysis (OLCA; Croon, 1990, 2002). The combination of these two models will allow researchers and practitioners to model student proficiency from an *explanatory* perspective (De Boeck & Wilson, 2004) expressed through the LLTM part of the model, while providing simple and interpretable results in terms of ranked performance groups, through the OLCA part of the model.

Section 2 of this paper briefly reviews the use of ordered performance levels in educational assessment contexts. Section 3 introduces ordered latent class analysis as a way of modeling ordered levels. Section 4 discusses the use of theory-based item features as a way of understanding the factors that govern the difficulty of tasks and how these are modeled in the LLTM model and the mixture LLTM model. Section 5 introduces and describes the OM-LLTM. Section 6 presents the results of a simulation study examining the performance of the model under different conditions, and Section 7 concludes with a discussion of the simulation results.

4.2 Ordered Performance Levels

The idea of interpreting test results in terms of ordered levels is not new, especially considering that in its most basic form, using tests to make pass/fail decisions implies the interpretation of results in terms of two ordered levels: one that fulfills a set of desired requirements and one that does not.

The definition of the levels has traditionally been with the assumption that a qualitative distinction can be made on the basis of different ranges of an underlying quantitative attribute. From this perspective, the definition of ordered levels has been treated in the psychometrics literature as a matter of defining one or more cut-points that will segment the quantitative attribute (Zieky, 1995). As such, the definition and determination of ordered levels has been extensively discussed as part of the literature on the procedures for establishing such cut-points, commonly known as *standard setting* methods (Cizek, 2001). In other words, the focus on ordered performance levels has been traditionally on the procedures for agreeing on cut-points, in order to

assign score ranges to descriptions that characterizes what the students in that range should be able to do, know or understand.

Setting cut-points—and in doing so defining a set of ordered levels— can play three different roles according to van der Linden (1995): (a) they define each level in terms of the kinds of problems that respondents can be expected to solve, (b) they simplify the presentation of results, and (c) they define "targets for the outcomes of educational policies in terms of the achievements of the examinees" (p. 98).

Indeed, ordered levels are a common tool used in educational assessments to characterize student performance. Labels such as "below basic," "basic," "proficient," "advanced," are common examples of a set of ordered performance levels that can be found in educational settings.

The practice of using these kinds of levels and labels has become increasingly common in the United States since the No Child Left Behind Act of 2001 (NCLB; Public Law 107-110) established the categorization of student performance in such terms as legal requirements. Additional examples of this practice can be found across states (Perie, 2008), in the NAEP achievement levels (Bourque, 2009), and as a way of communicating outcomes in international tests like PISA (OECD, 2007) and TIMSS (National Center for Education Statistics, 2009).

As is the case with many of the aforementioned examples, the construction of performance levels often arises as a product of practical goals or necessities. In line with van der Linden's (1995) set of three goals, ordered levels have been used to define and monitor legally required target outcomes (as is the case with the NCLB act), and to both define performance levels in terms of the problems students can solve and as a way of communicating assessment results to stakeholders with non-technical backgrounds:

> To help users interpret what student scores mean in substantive terms, the scale is divided into proficiency levels.... The levels range from the lowest, Level 1, to the highest, Level 6. Descriptions of each of these levels have been generated, based on the framework-related cognitive demands imposed by tasks that are located within each level, to describe the kinds of knowledge and skills needed to successfully complete those tasks, and which can then be used as characterisations of the substantive meaning of each level. (OECD, 2014, p. 46)

As this quote hints at, performance levels need not solely arise out of practical concerns, as they can also be developed based on theories of cognition and instruction (Mislevy, 1996; National Research Council, 2001; Wilson, 2005), representing hypotheses about the way in which students' understanding of a domain develops, such as the recent movement that promotes the characterization of developmental pathways in terms of *Learning Progressions* (LP; Lehrer & Wilson, 2011; National Research Council, 2006, 2007; C. Smith et al., 2006; Wilson, 2009).

The goal of connecting assessment and cognitive theory can be traced back at least to the 1960's with the introduction of criterion-referenced assessments (Glaser, 1963), and has gained considerable visibility in the past decade with the emergence of reports from the National Research Council, such as *How People Learn* (National Research Council, 2000) and *Knowing what Students Know* (National Research Council, 2001). The former presented an overview of how advances in cognitive psychology were informing our understanding of student learning and discussed how to create better instructional environments and assessments based on that knowledge, while the latter discussed in detail the role of cognitive theory as a necessary component of the development of educational assessments.

Conceptualizing cognitive theories in terms of ordered performance levels offers several advantages, as it presents a coordinated framework for research, instruction and assessment, while at the same retaining all the benefits in terms of inference and communication of results common to the practically motivated performance levels previously mentioned. The use of performance levels offers a framework for expressing the results of assessments in a diagnostic manner to teachers, students and other stakeholders, facilitating the interpretation of results and sharing in this sense the advantages usually associated with criterion-referenced assessments.

4.2.1 Characterizing the levels

However, the potential advantages of ordered levels as a way of facilitating the interpretation of assessment results hinges on the quality of the characterization of each level. This point can be separated in terms of two issues: How we interpret what a *level* is and what features we use to describe each level.

The first issue, the interpretation of *level*, is discussed extensively in Section 4.3, where we examine whether these levels are intended to be interpreted as qualitatively

distinct groups or simply as parts of a continuum that is being separated by cut-points. After all, even though the procedure for estimating ordered levels often corresponds to the latter, it is unclear to what extent this is understood by the public that receives the results in the form of performance levels.

The second issue, the quality of the elements used to describe each level, is tackled in Section 4.4. This is a critical element, as descriptions of the levels may vary significantly in terms of the amount and type of detail used to describe them, but they play an important role in facilitating inferences regarding student capabilities. The descriptions associated with each level support the interpretation of the assessment results, summarizing inter-individual differences in terms of membership to these different levels. In Section 4.4 we discuss how we use theory-defined item properties to characterize each one of the performance levels.

4.3 What is a level? Modeling Ordered Levels of Proficiency

We can think of ordered levels in at least four ways, ranging from a quantitative structure to an ordinal structure, represented in Figure 4.1.

Starting in pane (a) of Figure 4.1 a set of ordered levels can be thought of as a discretized quantitative variable. Think for instance of the signs that determine a minimum height for children to take part in a park ride or the idea of minimum age for purchasing alcohol; these are two examples where we conceive of two different kinds of persons as a function of their position in relation to a single cut-point. In both cases we understand that each person can occupy potentially infinite locations in the quantitative variable, but for the purposes of getting on a ride or buying scotch, we consider them solely in terms of their membership in the class that is either below or above the cut-point. This is perhaps the most common way of thinking about ordered levels.

A second way of conceptualizing a set of ordered levels is presented in pane (b) as a set of located heterogeneous classes, where we think that each class has an associated location along a quantitative attribute, but assume that members of each class manifest variation around the overall location of their class (hence the heterogeneous part of

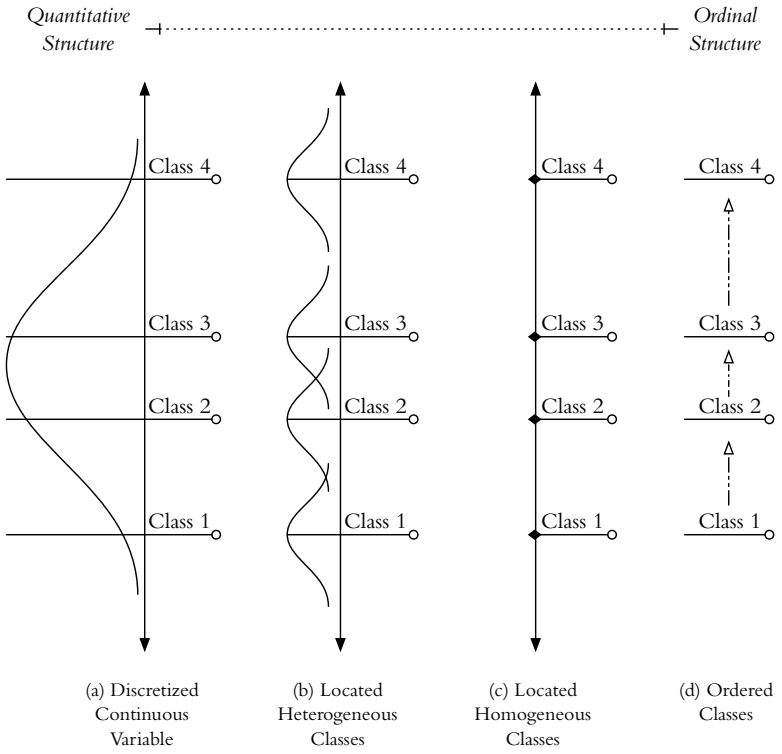


Figure 4.1: Representation of Alternative Models for Ordered Levels.

the name). This second kind of ordered level is less commonly considered, but this kind of ordered set can be seen for instance in the effects of an experiment with multiple conditions as a function of dosage of a medical treatment; each condition defines a level of dosage, and will produce a certain average effect on the outcome variable (assuming that the treatment has some kind of effect). This effect will rarely be homogeneous, presenting variability around an overall or average class location. Note that an important difference between the conceptualizations in panes (a) and (b) is that under pane (a) a member of a higher class, say class α, will always have a higher value on the underlying quantitative variable than a member of a lower class β; this is not the case under pane (b), where a member of the class α can potentially occupy a lower position than a member of class β.

A third option is to consider the ordered levels as homogeneous located classes as presented in pane (c). An easy way of thinking about this is to think of objects that have been constructed to a set of fixed specifications, for instance, the diagonal sizes

of the screens of different smartphone models. The idea here is that we think of the different classes do not have any variation within them, but still being clearly located on a quantitative underlying variable. Another way of thinking about homogeneous located classes is in relation to heterogeneous located classes, where the homogeneous class is the ideal specification for the constructions of a set of different objects, for instance the lengths of different kinds of screws, and the heterogeneous located classes represent the actual outcomes such as the actual length that come out of the process.

The fourth and last option considered here is represented in pane (d) as ordered classes. The key point under this way of thinking of classes is that there is no longer an assumption of an underlying quantitative variable. Examples of ordered scales are for instance the Mohs scale of mineral hardness (Mohs, 1825), the Beaufort wind force scale (Simpson, 1906), and any number of pain self report scales (Jensen & Karoly, 1992) which ask patients to rank their pain selecting a number from 1 to N. A common reaction to the idea of ordered classes in this sense is that *there must be an underlying quantitative variable*, even though mathematically, an ordered relation does not require an underlying quantitative structure; according to the formalization of quantity (Hölder, 1901; Huntington, 1902) order is a prerequisite of quantity, not the other way around (Michell, 1990). When we conceptualize the ordered levels in the forms represented in panes (a), (b) and (c) in Figure 4.1 we are tacitly indicating that all the differences between the classes can be summarized by a single distance between the locations of the classes in the underlying quantity. This is not the case for the ordered classes represented in pane (c), where we can simply assert an order relation, attributable to many factors and therefore not reducible to a single distance on a quantity.

These four ways of thinking about ordered levels can be modeled using multiple statistical approaches, which will be briefly described in the following sections.

4.3.1 Discretized continuous variables

Generally speaking, the use of a discretized continuous variable (Glass, 1978; Wright & Stone, 1979) relies on a two step process. The first part usually corresponds to the estimation of a continuous latent variable model, while the second part is the aforementioned 'standard setting' procedure (Cizek, 2001), used to divide the original continuous scale, as represented in Figure 4.1a.

Although of the four ways considered here this is arguably the most problematic, it is currently the most common approach. Conceptually, there is a discordance between the way in which the latent variable is being modeled (i.e. as a continuous distribution) and the classification indicated by the theory; the main consequence of this disconnection between the theory and the statistical model is the fact that the class membership of the respondents is not a structural part of the model. From a practical perspective, the need to conduct a separate procedure for determining the membership raises considerable issues related to the composition and legitimacy of the teams that will decide the cut-points as well as the validity of their decisions.

4.3.2 Located latent classes

The second approach, the use of heterogeneous located latent classes, has been studied as a special case in the framework of Factor Mixture Analysis (FMA; Lubke & Muthen, 2005; Titterington, Smith, & Makov, 1985), as an extension of both Latent Class Analysis (LCA Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968) and Factor Analysis (FA; Bartholomew & Knott, 1999). Applications of FMA have been described by Muthen and Asparouhov (2006) as a form of IRT mixture modeling, and these applications allow for the estimation of a model that includes multiple classes on a continuous factor with varying means and variances, as illustrated in Figure 4.1b.

The third alternative, the use of homogeneous located latent classes, was originally introduced by Lazarsfeld and Henry (1968) and has been later developed by Formann and Thomas (2002) and Uebersax (1993) as a combination of LCA and IRT. These kinds of models assume that all the respondents that belong to a particular class share the same level of proficiency. Therefore, they can be represented by a single location on the continuous scale, as illustrated in Figure 4.1c. It is worth mentioning that the latter can also be seen as a special case of FMA, where the variance of the located latent classes has been set to zero, an approach that does not assume a particular distributional shape of the continuous latent variable(Lindsay et al., 1991; Muthen & Asparouhov, 2006).

Up to this point, all the approaches have been based on a continuous representation of the latent variable of interest. This is consistent with a traditional IRT perspective where the variable being assessed, mathematical proficiency for example, is a

continuous variable in which the inter-individual differences between low and high performing students are a matter of degree.

However, the way in which ordered levels are conceptualized and described does not always coincide with this underlying assumption, because they rely on a set of levels that, although ordered in terms of proficiency, represent qualitative differences in understanding, not just differences of degree.

4.3.3 Ordered latent classes

It is in light of the issue of potential mismatch between a model of qualitatively distinct levels versus a model of a quantitative variable that the LP framework appears ideally suited to be explored using LCA, and more specifically, its extension to Ordered Latent Classes Analysis (OLCA; Croon, 1990, 2002).

As was the case with the use of homogeneous latent classes, under OLCA all the members of a single class have the same response probability of correctly answering the items; however, under OLCA no class location is estimated because the differences between the classes are assumed to be ordinal. This is achieved through the application of inequality restrictions to an otherwise typical LCA analysis (Croon, 1990, 2002), such that "one may say that the probability of a 'positive' response increases [for all items] when one runs through the set of latent classes from the 'lowest' to the 'highest' one" (Croon, 2002, p. 140). Ordered Latent Class Analysis has the added benefit that it combines the strengths of non-parametric approaches with a model based approach (Croon, 2002; Hojtink & Molenaar, 1997; Van Onna, 2004).

Therefore, the analysis of a LP through OLCA offers an alternative conceptualization of the nature of learning to the perspectives underlying two of the most prominent approaches currently adopted by the psychometric community, namely, Item Response Theory (IRT; Hambleton, Swaminathan, & Rogers, 1991) and Cognitive Diagnostic Models (CDM; Rupp et al., 2010), which rely either on a highly specific set of skills that can be characterized as being present or absent in the student(i.e. mastered or not mastered) in the case of CDM, or as a reduced number of quantitative variables relies on representing different levels of proficiency as (usually unidimensional) distances. Learning progressions and OLCA advance a distinctive understanding where each level represent an overall state of understanding of a domain, presenting in a sense a middle

ground between the continuous variables used in IRT and the multitude of specific skills that are commonly described under a CDM perspective.

Generally speaking, a mixture model defines the probability of a given response pattern for a respondent conditional on the class to which the respondent belongs (Heinen, 1996), or in this case, the strategy (or sets of strategies) that the respondent utilized:

$$\Pr(\mathbf{x}_p) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} \left(\pi_{i|c}^{x_{ip}} \times (1 - \pi_{i|c})^{1-x_{ip}} \right) , \tag{4.1}$$

where persons are indexed by p from 1 to P, tasks are indexed by i from 1 to I, strategies are indexed by c from 1 to C, and

π_c is the probability that a person belongs to class c, and $\sum_{c=1}^{C} \pi_c = 1$.

$\pi_{i|c}$ is the conditional probability of a person in class c presenting a correct answer x_i to task i given that they belong to class c.

Equation 4.2 shows that the logit of the conditional probability is specified as an interaction parameter β_{ic} between class c and task i.

$$\text{logit}[\pi_{i|c}] = \beta_{ic} \tag{4.2}$$

The traditional LCA is unconstrained, which means that the $\pi_{i|c}$ are free parameters, allowing a class-specific response probability for each item. This flexibility means that the classes are qualitatively different and cannot be directly compared under a metric. Yet, it is important to note that this flexibility comes at the price of a different, and arguably equally stronger, assumption about the respondents: the homogeneity of the respondents within each class.

Marcel Croon (1990, 1991) introduced ordered latent class analysis by applying inequality constraints, forming a simple order of the classes (Croon, 2002). Restricting the β_{ic} parameters allows interpretation of the classes in terms of an overall ranking of proficiency, where a lower c implies a lower level of proficiency, as illustrated in Equation 4.3.

$$\begin{aligned} \text{logit}[\Pr(x_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] = \beta_{ic}, \\ \beta_{ic} \leq \beta_{ic'} \text{ for all } c < c' \text{ and for all } i. \end{aligned} \tag{4.3}$$

In this way, OLCA offers an interpretation similar to the one used to interpret traditional IRT models in terms of placing respondents in a hierarchy of proficiencies, but this is done without assuming that differences between two classes can be summarized solely by a distance in a single quantitative variable.

The previous section raised two questions regarding the characterization of ordered levels: how do we interpret what a *level* is? and what features do we use to describe each level? This section has tackled the first of this questions, discussing four alternative conceptualizations and alternative statistical approaches that can be used to model each one of them, but we have focused in particular in the conceptualization of ordered levels as ordered classes and the ordered latent class models that can be used to model them, which will be incorporated into the OM-LLTM. The second question regarding the elements that we can used to more meaningfully describe each levels is discussed in the following section.

4.4 Modeling item features: the LLTM and mixture LLTM

Having discussed the issue of the conceptualization of levels, we now turn to the question of what features we can use to describe each level. From a summative perspective the classification of students may be enough; after all, there are conceivable applications where ranking is all that might be needed (e.g. selection of applicants for a reduced set of vacancies). However, it is more often the case that we want more information than that from our assessments; we want to interpret and communicate the results in a meaningful way, and that requires understanding what membership of each ordered class means.

The bare minimum to characterize a set of ordered classes is to assign a label to each of them, generally communicating whether belonging to them is supposed

to be considered something positive or negative, say, 'high performance,' 'expected performance,' and 'low performance.' Beyond labels, we can add general descriptions of what is expected of each one of these levels, either as a function of prior theory that indicates what these levels should be or empirically based characterization rooted in the expected performance of each one of the levels on the tasks that were part of the assessment. In both cases, the descriptions help us interpret what it means to be assigned to each of the levels. However, general performance descriptions, while more helpful than mere labels, are not necessarily detailed enough to provide diagnostic information about the different ordered levels.

This is where the key idea of this section comes in: we can characterize the levels in a diagnostic manner by taking advantage of a close connection between our theories about the domain we are assessing and the mathematical models we use to analyze them in order to leverage more meaningful interpretations. This can be achieved by modeling item features that—according to our theory—govern how challenging the items are to respondents. What is the advantage of doing so? That we can now characterize each of the ordered levels in terms of how well people at these levels are perform in these key features that determine the difficulty of the tasks in the domain that is being assessed.

4.4.1 Embedding Theory into Items: Item Features

Let us start with a simple example: suppose we are modeling performance in an early arithmetic test. We could have a very simple hypothesis that the difficulty of the items in the test is a function of the operations involved. What does this mean? It means that as a part of our very simple hypothesis, we think that it is reasonable to expect that all of our 'addition' items will have a similar difficulty and that all the 'multiplication' items will also have a similar difficulty, and so on. Based on this assumption we could model the difficulty of all the items in this test as a function of just four features: summation, subtraction, multiplication, and division.

In this way, instead of estimating the difficulty of each test item, we would estimate the difficulty of solving an 'addition' item, a 'subtraction' item, a 'multiplication' item, and a 'division' item, and we could interpret the person location or level of performance in terms of their likelihood of having mastered each one of these features, as is represented in Figure 4.2.

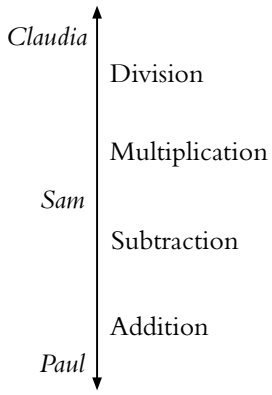


Figure 4.2: A Diagram of the Locations of Items (as Function of Features) and Persons.

With this simple model we can potentially interpret the results of our test in direct relation to the features that we are interested in, learning that while Claudia seems to have no problem solving all of the operations, Sam seems to have learned enough to solve only addition and subtraction problems, and Paul has yet to learn how to solve any of the four operations.

This is of course an extremely simplified example, but it highlights how connecting items directly to theoretical features can help us obtain more meaningful and diagnostic interpretations of an assessment. This idea was introduced into psychometrics in 1973 by Gerhard Fischer (1973) in his paper introducing the Linear Logistic Test Model (LLTM), a model that formalizes this idea of modeling tasks in terms of a set of relevant features.

4.4.2 The Linear Logistic Test Model

The Linear Logistic Test Model (LLTM; Fischer, 1973), is a powerful tool for investigating whether the difficulty of a set of tasks can be modeled as a function of a set of theoretically defined features, which can be interpreted as discrete cognitive operations (Formann, 1995).

The LLTM was developed as an extension to the Rasch Model (RM Rasch, 1960/1980), which modeled the probability of successfully responding to a task as a trade-off between the proficiency of the respondent and the difficulty of the task. This basic proposition was formalized by Georg Rasch as presented in equation 4.4 under a latent variable framework.

$$\text{logit}\left[\Pr(x_{pi} = 1|\theta_p)\right] = \theta_p - \delta_i \tag{4.4}$$
$$\theta_p \overset{iid}{\sim} N(0, \psi)$$

Where:

x_{pi} represents the response of person p to task i.
θ_p stands for the person p's proficiency.
δ_i indicates the difficulty of task i.
ψ represents the variance of the person distribution.

The LLTM extends the RM by decomposing the item difficulty into a set of basic features, such that instead of specifying one difficulty parameter for each task, the model structures the difficulties as a linear function of the smaller set of features that characterize the tasks. At this point it is important to note that the use of the LLTM demands a clear theory that justifies the decomposition of tasks into basic features, a non trivial requirement in comparison to the requirements for the RM and most other IRT models as well.

Following from our previous simple example, this idea is illustrated in Table 4.1, where the table indicates how six arithmetic tasks can be decomposed into the four basic operations, as represented by the first four columns in Table 4.1. The decomposition of the tasks can also be made in terms of interactions of factors, as illustrated in the last two columns in Table 4.1, where the second to last column is indicating the presence of more than one operation, and the last column is indicating the simultaneous presence of both multiplication and division in the item. Notice that, as previously noted, we consider the difficulty of the second and third tasks to be the same, because both of them are characterized by the same feature, in this case 'subtraction'.

This example also shows that in an LLTM each task is associated with a vector of values, a row in Table 4.1, that indicates which features are present in each task. Formally, this vector q_i allows us to express the δ_i parameters as a linear combination of K basic features as seen in equation 4.5.

	Summ.	Subt.	Mult.	Div.	N Ops. > 1	$\times$ & $\div$
$8 + 3$	1	0	0	0	0	0
$2 + 2$	1	0	0	0	0	0
$5 - 1$	0	1	0	0	0	0
$8 - 3$	0	1	0	0	0	0
7×2	0	0	1	0	0	0
$9 \div 3$	0	0	0	1	0	0
$(9 - 2) \times 3$	0	1	1	0	1	0
$(3 \times 8) \div 4$	0	0	1	1	1	1
$(5 \times 3) + (8 \div 2)$	1	0	1	1	1	1

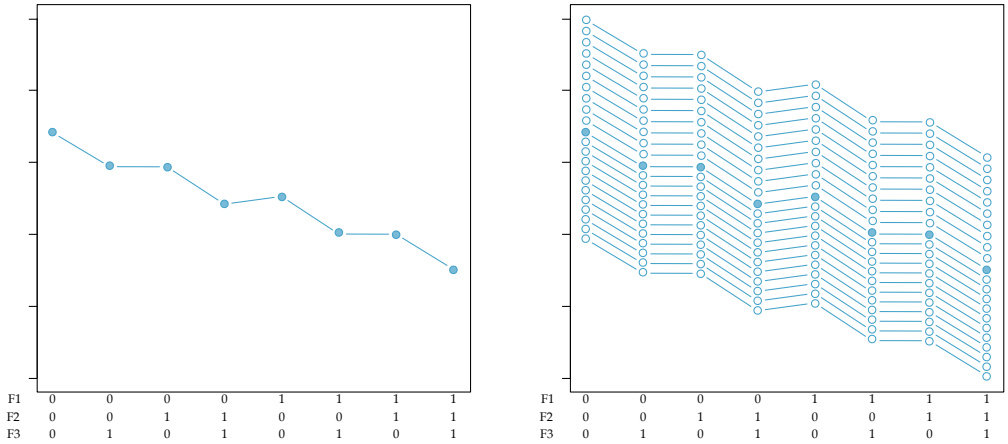
Table 4.1: Set of Items as a Function of a Matrix of Features.

$$\text{logit}\left[\Pr(x_{pi} = 1|\theta_p)\right] = \theta_p - \left(\eta_0 + \sum_{1}^{K} \eta_k q_{ik}\right) \tag{4.5}$$
$$\theta_p \overset{iid}{\sim} N(0, \psi)$$

This model formalizes our hypothesis about which of these features play a role in determining the difficulty of the tasks, and the extent of the effect of each k feature is represented by the η_k parameters.

By modeling the results of an assessment using the LLTM what we obtain is an overall set of weights associated with the difficulty of each feature. We can represent these weights as a logit pattern as in Figure 4.3a, which plots a set of mock results for a test with three dichotomous features. In this Figure, which we can call a Response Pattern Plot, the x-axis shows the eight possible combinations of three dichotomous item features, while the y-axis shows the log odds or logit associated with correctly answering an item with those combinations of attributes.

As mentioned before, the pattern in Figure 4.3a represents the overall weights estimated, but the actual logit associated with each person will vary around these overall weights as a function of the random effect θ_p associated to each person. Therefore, if we wanted to diagram the patterns for the respondents, we would get something sim-



(a) Overall pattern

(b) Patterns for the LLTM

Figure 4.3: Logit Response Patterns

ilar to Figure 4.3b, where the patterns for individual respondents will be translations of the overall pattern as a function of their estimated random effect.

A key point to highlight here is that the LLTM presupposes that the item features present the same pattern of challenge to every respondent, and the only difference between respondents is whether this pattern is translated downwards (i.e. less likely to respond correctly) or upwards (i.e. more likely to respond correctly) depending on the person specific θ_p. However, it is not unreasonable to think that the different item features may present different challenges to different students depending on different kinds of approaches or solving strategies that they could bring to the tasks; what may be an easy feature for some could be a hard feature for others. Effectively, this would mean that different kinds of students could present very different patterns, no longer being simply translations of a single overarching one, which would require an extension to the LLTM that would allow us to (a) detect different kinds of patterns, and (b) would tell us which respondents are likely to be associated with each one of those patterns, a task ideally suited for a latent class or mixture model.

4.4.3 The Mixture Linear Logistic Test Model

As Mislevy and Verhelst (1990) note, traditional IRT models characterize respondents in terms of an *"overall propensity towards correctness"* (p. 195), which is consistent with the assumption that all the respondents approach the task in the same manner. In other words, the difficulty of a task is not absolute in a given domain, but contingent on the strategy that is adopted to solve it.

Accordingly, the respondent proficiency of these models should not be considered directly as their proficiency in the domain in general, but should be interpreted as the respondent's success when approaching the domain of interest under a specific strategy (Mislevy & Verhelst, 1990), namely, the predominant strategy used by the test respondents.

To address this issue, Mislevy and Verhelst (1990) introduced a mixture version of the LLTM. This model allows the estimation of a set of proficiencies for each respondent, each based on a different strategy, and the probability that the respondent used each one of these strategies. It is important to note however, that this model requires the definition, in advance, of the number of strategies that will be examined, further increasing the demands on the theory that is required to analyze a test ; a requirement that is somewhat alleviated by the possibility of empirically examining how many classes can be recovered by comparing variations of the models with increasingly more classes and comparing their relative fit.

A mixture model, as presented in Equation 4.1, is composed both of a set of parameters π_c that model the proportion of respondents that belong to each class c and a set of $\pi_{i|c}$ parameters that model the response probability of person p to item i conditional on that person belonging to class c.

In the case of the mixture LLTM, this last component, $\pi_{i|c}$ corresponds to an LLTM model for strategy c as shown in equation 4.6.

$$\text{logit}[\pi_{i|c}] = \theta_{pc} - \sum_{k=1}^{K} \eta_{kc} q_{ik} \tag{4.6}$$

such that the M-LLTM models a response vector as follows:

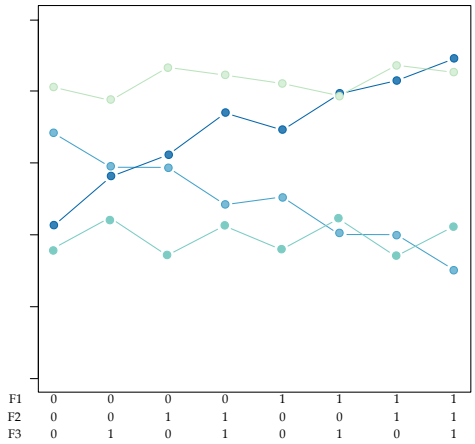


Figure 4.4: Diagram of the Overall Patterns Under an M-LLTM.

$$\Pr(\mathbf{x}_p|\boldsymbol{\theta}_p) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} \left[\text{logit}^{-1}\left(\theta_{pc} - \sum_{k=1}^{K} \eta_{kc} q_{ik}\right)^{x_{ip}} \times \left(1 - \text{logit}^{-1}\left(\theta_{pc} - \sum_{k=1}^{K} \eta_{kc} q_{ik}\right)\right)^{1-x_{ip}} \right] \tag{4.7}$$

$$\theta_{pc} \overset{iid}{\sim} N(\mu_c, \psi_c) \; .$$

By allowing the weight of the features to vary by classes through the η_{kc} parameters, the mixture LLTM allows the difficulty of the basic task features to depend on the strategy that is being utilized. Thanks to this flexibility, the M-LLTM is no longer modeling a single overall pattern for the entire set of respondents, but modeling C different patterns, as illustrated in Figure 4.4.

The M-LLTM is effectively modeling multiple LLTM models, one for each of the C strategies that we expect to find, and it allows us to estimate the probability that each respondent has of being associated with each one of these strategies. However, the flexibility of the M-LLTM comes at the cost of a considerably more complex interpretation of the results: now every respondent has not one, but a vector of C proficiencies

, one for each strategy. In a sense, the M-LLTM is providing us with potentially useful diagnostic information by telling us what is the most likely strategy or approach that each student is using, but at the same time, the model seems incompatible with most educational contexts that require some kind of decision making. After all, we now have not one but multiple scores for each person, corresponding to a set of strategies that are not necessarily comparable or ordered. One could consider then just assigning to each person the θ_{pc} estimate associated with the class c that the person is most likely to belong to, however, this would result in an uninterpretable comparison of proficiency estimates under different strategies, not a comparison of proficiencies under a single or overall strategy.

Ideally, what we want is a model that (a) is capable of distinguishing different kinds of respondents in terms of their different "strategies" (a strength of mixture models), (b) can decompose item difficulty in terms of the item features defined by our theory (the contribution of the LLTM) but that, at the same time, (c) offers us an interpretation in terms of a single result (i.e. the most likely class membership) that estimates—and allows us to communicate—the overall performance of the respondent, a feature of ordered latent class models.

4.5 The Ordered Mixture Linear Logistic Test Model

The OM-LLTM assumes respondents are grouped in *ordered* latent classes where the probability of correctly answering an assessment task is a function not only of the class membership of the respondent, but also of item features that—according to the theory—determine the difficulty of the task. Unlike the M-LLTM the OM-LLTM does not rely on the estimation of a quantitative latent variable θ, providing a general ranking of respondents based on class membership, instead of relying on a class specific θ.

The OM-LLTM is defined using the same set of inequality constraints as the OLCA models. The only difference is that instead of estimating I unique item parameters for each class, we summarize the I parameters into K attribute parameters using a q_i vector, just as in the LLTM:

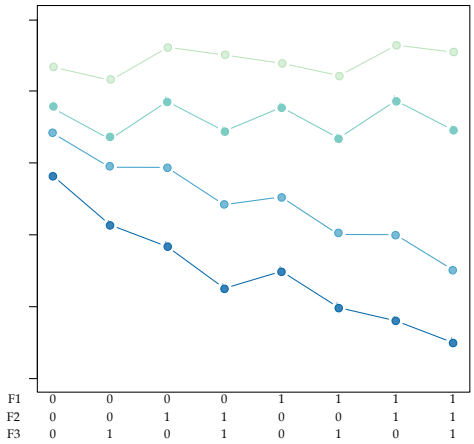


Figure 4.5: Diagram of the Overall Patterns Under an OM-LLTM.

$$\text{logit}[\Pr(x_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] = \eta_{0c} + \sum_{k=1}^{K} \eta_{kc} q_{ik} = \beta_{ic}, \tag{4.8}$$

$$\beta_{ic} \leq \beta_{ic'} \text{ for all } c < c' \text{ and for all } i.$$

In this way, the OM-LLTM allows us to estimate differential patterns that are at the same time constrained to be ordered, as illustrated in Figure 4.5.

We now have enough flexibility to detect differential approaches (reflected in different classes and their corresponding patterns), but at the same time we have organized those patterns in a hierarchical relation, such that membership to a certain class can readily be interpreted in terms of overall proficiency, which was the issue under the M-LLTM,which relied on to the inclusion of a set of class-specific latent variables θ_{pc} that were not directly comparable.

As one potentially interesting variation to the model we can place the constraints directly on the η parameters. Where the η parameters can be interpreted as the increase in "easiness" when the item feature is present compared with when it is absent within each ordered class; such that the least proficient classes will have smaller η, i.e. they will be less likely to correctly answer the items that exhibit those proper-

ties, while the more proficient classes will have larger η, increasing the probability of answering correctly when the properties are present.

If the constraints are applied to all the η parameters, then we are effectively ordering according to all the effects of item properties :

$$\text{logit}[Pr(x_{ic} = 1|c)] = \text{logit}[\pi_{i|c}] = \eta_{0c} + \sum_{k=1}^{K} \eta_{kc} q_{ik}, \tag{4.9}$$
$$\eta_{kc} \leq \eta_{kc'} \text{ for all } c < c' \text{ and for all } k.$$

This last variation opens up the interesting possibility of only applying the constraint to some of the η parameters. Effectively, this would allow one to order the classes according to certain attributes that may be deemed more important (e.g. the focus of a given curriculum), while allowing the rest to vary freely.

4.6 Simulation Study

In order to examine parameter recovery for the OM-LLTM performs under different conditions, a simulation study was conducted examining the impact of the number of respondents, the number of tasks included in the assessment, and the number of item features.

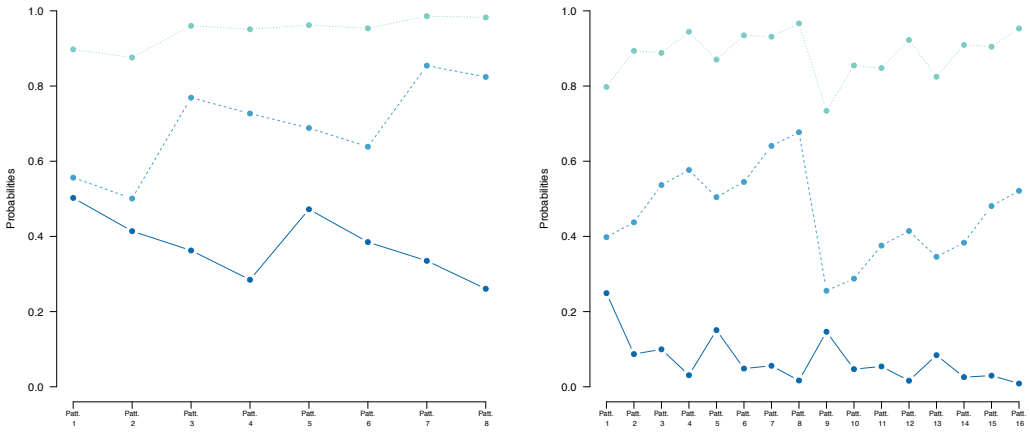
The first two factors, number of persons and number of tasks reflect two aspects that will likely vary under different application conditions, and are therefore worth examining to assess their impact on the recovery of the OM-LLTM parameters. The third factor, the number of features, considers the complexity of the theory in the characterization of the items, and is included to investigate how different levels of complexity impact recovery in conjunction with varying numbers of respondents and items.

The expectation is that more persons and items allow for better estimation of the model due to the additional information available, while additional features should put a strain on the estimation because additional parameters need to be estimated.

Each of these three factors were incorporated into the simulation using two levels: 500 vs 1000 respondents, 20 vs 60 tasks, and 3 vs 4 factors. This structure produces a simulation design with eight cells. 100 replications were performed per cell

For the purpose of this study, the number of ordered classes was fixed at three. The respondent distribution among the three classes was randomized under each replication using a multinomial distribution with a probability of 0.25 for the lowest class, 0.50 for the middle class, and 0.25 for the upper class.

The generating probability patterns and parameters were fixed, with one set of values for the simulations that included 3 features and a set of values for the simulations with 4 features. These features were used to generate 20 or 60 items. Each set of items contained one item assigned to each of the possible combinations of features (8 items for 3 factors and 16 items for 4 factors) and the remaining items randomly assigned from the possible featurefactor combinations. The generating probability patterns—i.e. the probabilities for each combination of features given by the inverse logit of equation 4.9—for the 3 and 4 features simulations are presented in Figure 4.6.



(a) Three features

(b) Four features

Figure 4.6: Generating Probability Response Patterns

The generation of the simulation data and the subsequent analysis and graphics of the results were conducted using R (R Core Team, 2015). The estimation of the OM-LLTM was conducted using the EM algorithm as implemented in LATENT GOLD (Vermunt & Magidson, 2005, 2008).

4.7 Results

To evaluate the results of the different simulation conditions I will focus on two main areas. The first one corresponds to the recovery of the membership of the respondents in the set of ordered classes and the second one is the recovery of the generating parameters that determine the recovery of the generating probability profiles and recovery of class membership.

4.7.1 Class Recovery

A first element to examine is the recovery of overall class proportions, which for all the simulations corresponded to three classes with probabilities of 0.25, 0.5, and 0.25. The classification of cases was done in LATENT GOLD according to the maximum posterior probability:

$$\Pr(C_p = c|\mathbf{x}_p) = \frac{\widehat{\pi}_c \prod_{i=1}^{I} \left(\widehat{\pi}_{i|c}^{x_{ip}} \times (1 - \widehat{\pi}_{i|c})^{1-x_{ip}}\right)}{\sum_{c=1}^{C} \widehat{\pi}_c \prod_{i=1}^{I} \left(\widehat{\pi}_{i|c}^{x_{ip}} \times (1 - \widehat{\pi}_{i|c})^{1-x_{ip}}\right)} \tag{4.10}$$

Figure 4.7 presents the results of the recovery of class proportions under the eight conditions.

The diagram shows that the estimated bias of π_c was small across the simulations was good in general in all the cell conditions. The conditions with more respondents and more tasks showed less variation than conditions with fewer respondents and tasks: it can be seen that the conditions with 500 persons and 20 items present a small bias for π_c. On the other hand, the number of features (3 versus 4 in this case) does not seem to significantly affect the recovery of the overall class proportions. However, regardless of the conditions, the bias was not statistically significant in under any of the conditions.

A second aspect that is important to examine beyond the overall class proportion recovery is the accurate assignment of respondents to classes under each one of the simulation conditions. Class assignment was done using Maximum A Posteriori as a criterion, effectively assigning the class that yielded the highest probability of observing that pattern as a function of both the class specific probabilities and the overall

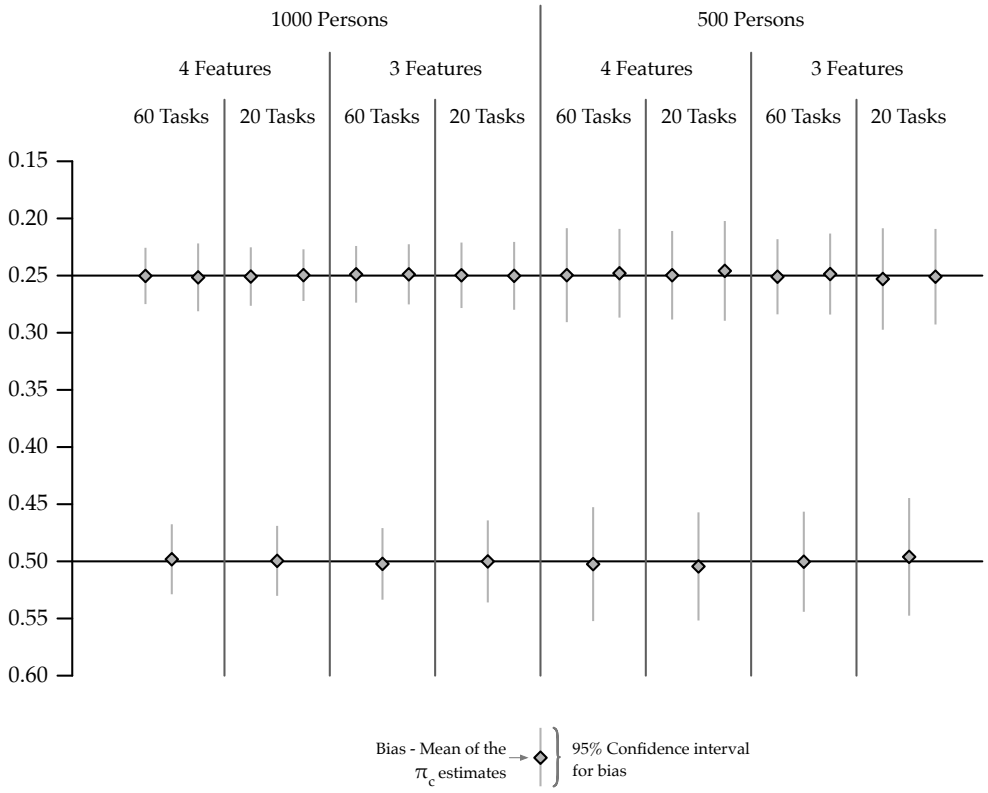


Figure 4.7: Class Proportions Recovered in the Simulation Study.

proportion of the classes.. The summary of the results for the person membership recovery is presented in Figure 4.8.

The results of the respondents class recovery among the eight cells in this simulation design ranged from a perfect recovery—under the condition with a 1000 respondents and 60 tasks and 4 features—to a good 82% under the condition with the lowest number of respondents and tasks and the higher number of features. The pattern of results are in general consistent with the expectation that a higher number of tasks and more respondents improve the estimation. Accordingly, the condition with 60 items uniformly recovered the correct membership for upwards of 99% of the respondents, and the only conditions with lower recovery rates involved the use of only 20 items.

	1000 Persons				500 Persons			
	4 Features		3 Features		4 Features		3 Features	
	60 Tasks	20 Tasks	60 Tasks	20 Tasks	60 Tasks	20 Tasks	60 Tasks	20 Tasks
Mean Proportion Correctly Recovered	1.0	0.99	> 0.99	0.94	> 0.99	0.82	> 0.99	0.94
Std. Dev. Proportion Correctly Recovered	0.0	< 0.01	< 0.01	< 0.01	< 0.01	0.27	< 0.01	0.01

Figure 4.8: Correct Membership Recovery in the Simulation Study.

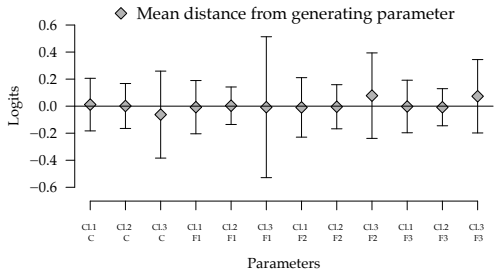
On the other hand, these results do not show a consistent pattern regarding the effects of adding extra parameters associated with the additional feature, as it can be seen that the recovery under the 500 respondents and 20 items condition with four features performed worse than the condition with only three features, but that was not the case for the conditions with 1000 respondents and 20 items.

4.7.2 Feature Coefficient Recovery

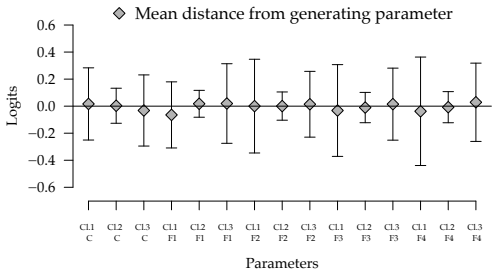
Regarding the recovery of the feature parameters, the four simulation conditions with 3 features had a total of 12 generating feature parameters (3 constants per class plus 3 feature coefficients per class), while the four simulation conditions with 4 features had a total of 15 generating parameters for the features (3 constants per class plus 4 feature coefficients per class). The summary of the simulation results is presented in Figure 4.9 on page 157; each one of the panes in Figure 4.9 presents a summary of the difference between the estimated parameter for each simulation and the generating parameter, such that if all the simulations had presented perfect recovery the means in the panes would be zero.

A first look at the results across all the panes in Figure 4.9 shows that the recovery of the feature parameters was in general good across all conditions, with little bias overall but differences in the variability of the recovered estimates.

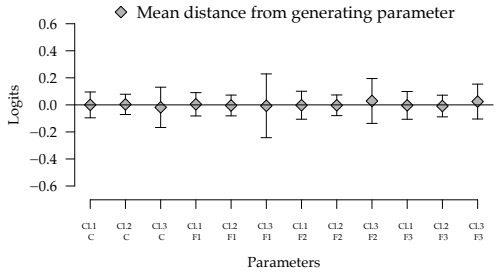
The left column of panes in Figure 4.9 shows the results for the conditions with 3 features, while the right hand column presents the conditions with 4 features; it is possible to see that for each one of the rows—which share the other two factors of



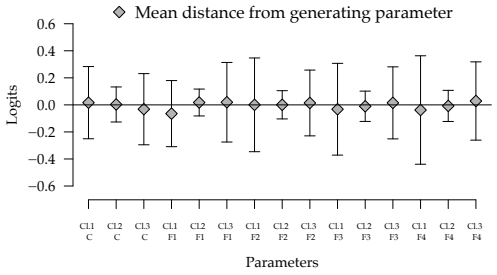
(a) Centered Means and 95% Conf. Int.
500 Respondents – 20 Tasks – 3 Features



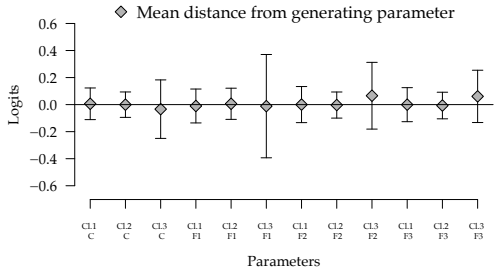
(b) Centered Means and 95% Conf. Int.
500 Respondents – 20 Tasks – 4 Features



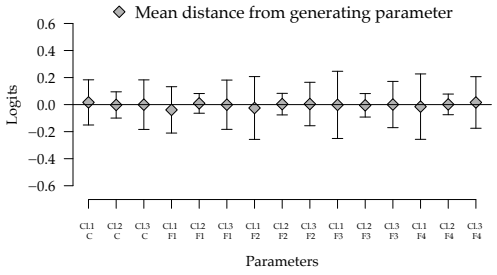
(c) Centered Means and 95% Conf. Int.
500 Respondents – 60 Tasks – 3 Features



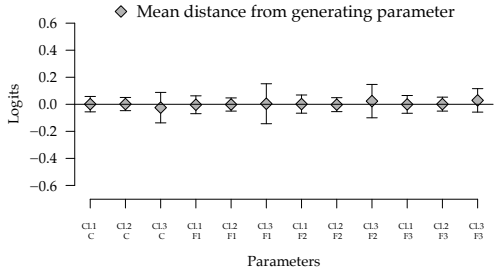
(d) Centered Means and 95% Conf. Int.
500 Respondents – 60 Tasks – 4 Features



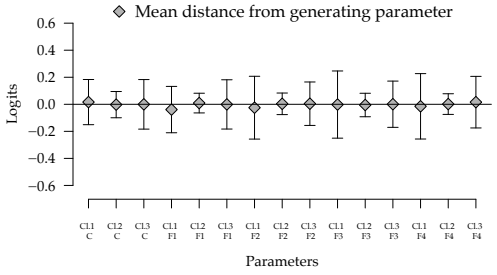
(e) Centered Means and 95% Conf. Int.
1000 Respondents – 20 Tasks – 3 Features



(f) Centered Means and 95% Conf. Int.
1000 Respondents – 20 Tasks – 4 Features



(g) Centered Means and 95% Conf. Int.
1000 Respondents – 60 Tasks – 3 Features



(h) Centered Means and 95% Conf. Int.
1000 Respondents – 60 Tasks – 4 Features

Figure 4.9: Recovery of Generating Feature Parameters

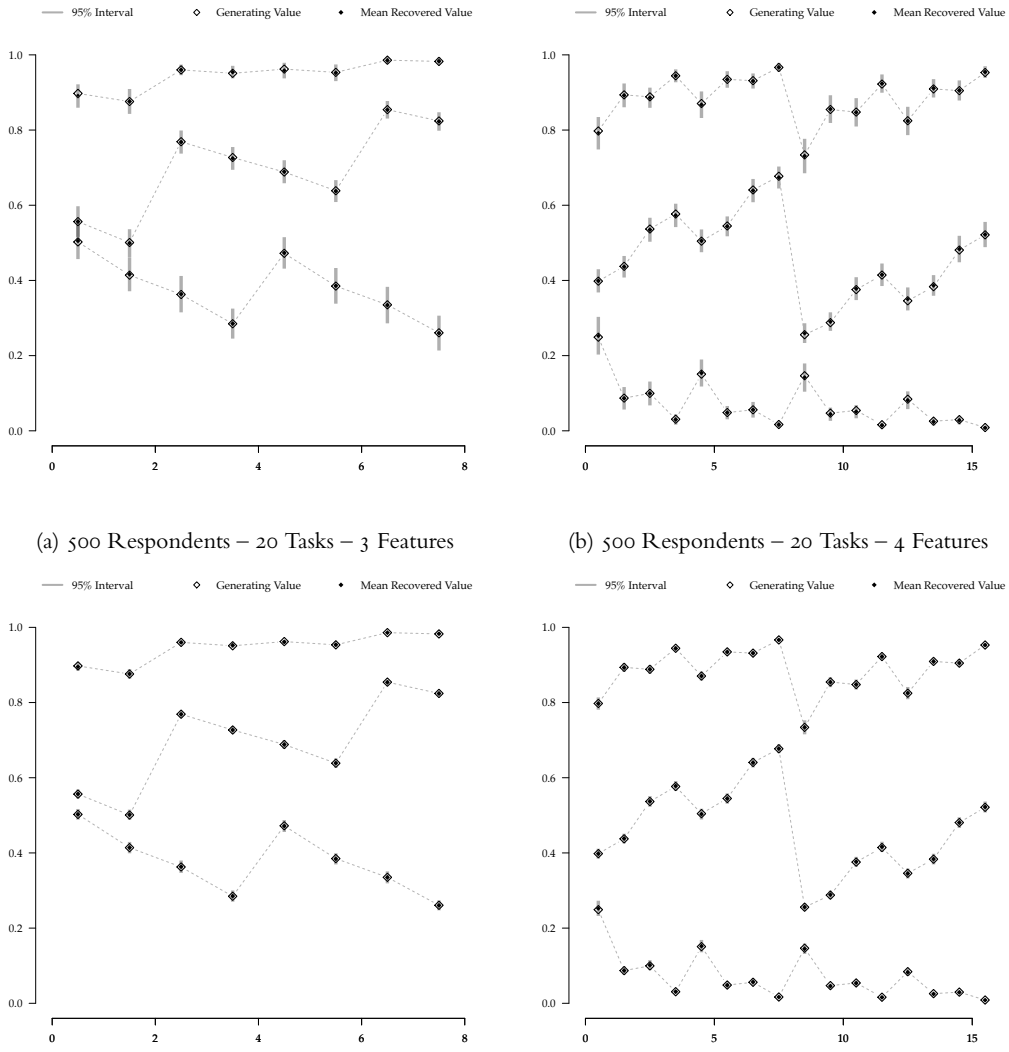
the design—the estimators tended to have a smaller variation for the conditions with 3 features. However, this is not uniformly the case, as can be seen for instance in pane 4.9a, where the feature parameter estimates of the third class were considerably more variable than the other two classes, making them similarly variable to estimates of the parameters in the condition with 4 features. Despite this, the overall results are consistent with the expectation that the conditions with additional parameters would, with the same number of items and persons, result in less precise estimates, which is reflected in the larger variability detected among the recovered parameters.

Regarding the number of tasks, rows 1 (panes 4.9a and 4.9b) and 3 (panes 4.9e and 4.9f) present the results for the conditions with 20 tasks, while rows 2 (panes 4.9c and 4.9d) and 4 (panes 4.9g and 4.9h) present the results for the conditions with 60 tasks. The comparison across the rows shows that, as it is reasonable to expect, the conditions with more tasks produced more precise results, expressed here as smaller variation in the parameter estimates.

Finally, comparing the upper half of Figure 4.9, which presents the results for the conditions with 500 respondents, versus the lower half also shows results consistent with the expectation that a more information provided as a function of higher number of respondents produced a lesser degree of variability among the parameter estimates.

In light of these tendencies, we can see that the best recovery can be found in panes 4.9g and 4.9h, while the most variable results are in panes 4.9a and 4.9b.

It is interesting to note that, regarding the different degrees of variability that can be seen within each one of the simulation conditions, there seems to be a systematic "class effect" rather than a systematic "feature effect." In other words, the results within the conditions show that the parameters associated with certain classes are better estimated across all features, and not that some features are better estimated regardless of class. One potential factor that may be related to this is the size of the class (as a function of the overall class proportion), where larger class could be better estimated. This hypothesis is supported by patterns in the right panes of Figure 4.9, namely the conditions with 4 features, where the middle class (with a generating proportion of .5) is consistently better estimated than the other two classes. However, this pattern does not hold for the panes on the left side of Figure 4.9, the conditions with 3 features, where the first two classes are estimated similarly well, while it is the third class that consistently shows more variability.



(a) 500 Respondents – 20 Tasks – 3 Features

(b) 500 Respondents – 20 Tasks – 4 Features

(c) 1000 Respondents – 60 Tasks – 3 Features

(d) 1000 Respondents – 60 Tasks – 4 Features

Figure 4.10: Probability Response Patterns

The results presented so far have shown the recovery of the feature generating parameters in the logit scale, but an alternative way to examine the results is to examine the results in terms of the variability that is exhibited in the recovery of the generating probability for feature patterns. Figure 4.10 presents the results for four of the eight conditions in terms of the probability for feature patterns, contrasting the variability under the conditions that performed poorly, relatively speaking, and

the best performing conditions, allowing us to contrast the results on the probability scale and interpret the variabilityprobability patternsin this terms.

Panes 4.10a and 4.10b show the conditions where the variability of the estimates was the largest, and it is possible to see the 95% interval around the mean estimated probability being, for instance, consistently greater in the bottom class in pane 4.10a. In these two panes we can see that at its maximum, the variability of the estimates represented by the 95% interval around the mean of the estimated probabilities reaches a range of approximately 0.1 in the probability scale.

On the other hand, in panes 4.10c and 4.10d the variability is considerably less, with the interval bars almost completely obscured by the means and generating parameter symbols. In these panes the largest 95% interval reaches 0.03 in pane 4.10c and 0.04 in pane 4.10d.

4.8 Discussion

This paper presents the OM-LLTM as a model that can draw on the strength of both the OLCA and the LLTM tradition to produce both overall simple interpretations while simultaneously examining the results in a more fine-grained level based on the underlying theory used to construct the assessment tasks. The simulation study presented here shows that the parameters of the OM-LLTM can be successfully recovered using the EM algorithm under different conditions using the off-the-shelves statistical package LATENT GOLD statistical package (Vermunt & Magidson, 2005, 2008).

The simulation conditions examined here are just an initial step, covering key factors that will affect the quality of the estimates, but by no means exhausting the potentially relevant factors. Future examination should consider, for instance, the impact of varying the numbers of classes, the introduction of features with more than two levels, and the overall separation of the classes, as this last factor will undoubtedly affect the accuracy of classification.

Areas of future development are the extension of this model to polytomous responses, the estimation of a model variation that places the constraints on the β parameters (allowing the η parameters to vary freely). Yet another potential extension is the use of partially ordered levels, allowing even more flexibility by modeling sets of classes that are ordered at the level of sets of classes, but only qualitatively different

within each set. Finally, an already mentioned potential variation of the OM-LLTM is the use of constraints on only some of the η parameters, allowing the ordering of classes only on features that are deemed relevant, while allowing other features to vary freely.

In sum, this paper makes the case that the OM-LLTM is a promising alternative to include in the measurement model both summative and formative concerns, and presents evidence from a simulation study that shows that the OM-LLTM can be successfully recovered and used to correctly assign class membership to respondents under a variety of conditions.

Bibliography

Adams, E. W. (1965). Elements of a Theory of Inexact Measurement. *Philosophy of Science*, *32*(3/4), 205–228. (Cit. on p. 37).

Adams, E. W. (1966). On the nature and purpose of measurement. *Synthese*, *16*(2), 125–169. (Cit. on pp. 36, 69–72, 78, 86).

Adams, E. W. (1979). Measurement Theory. In P. Asquith & H. Kyburg (Eds.), *Current research in philosophy of science* (pp. 207–227). East Lansing, Michigan: Philosophy of Science Association. (Cit. on pp. 35, 37).

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. (Cit. on p. 96).

Aera, Apa, & Ncme. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association. (Cit. on p. 76).

Aera, Apa, & Ncme. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association. (Cit. on p. 76).

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley-Interscience. (Cit. on p. 100).

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester, Mass: Clark University Press. (Cit. on p. 60).

Andersen, E. B. (1988). Comparison of latent structure models. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 207–229). New York: Plenum Press. (Cit. on pp. 95, 99).

Aristotle. (2001). Categories. In R. McKeon (Ed.), *The basic works of Aristotle* (pp. 7–37). New York: Modern Library. (Cit. on p. 20).

Bacon, M. (2012). *Pragmatism: an introduction*. Malden, MA: Polity. (Cit. on pp. 1, 2, 12–14).

Bartholomew, D. J. (1987). *Latent variable models and factors analysis*. New York: Oxford University Press, Inc. (Cit. on p. 98).

Bartholomew, D. J. & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold. (Cit. on pp. 100, 139).

Berka, K. (1983). *Measurement: its concepts, theories, and problems*. Dordrecht: Reidel. (Cit. on pp. 19, 27, 28, 44, 82).

Bernstein, R. J. (1998). Community in the pragmatic tradition. In M. Dickstein (Ed.), *The revival of pragmatism: new essays on social thought, law, and culture* (pp. 141–156). Durham: Duke University Press. (Cit. on p. 13).

Bernstein, R. J. (2010). *The pragmatic turn*. Cambridge, UK; Malden, MA: Polity Press. (Cit. on pp. 10, 14, 15).

Binet, A. (1900). *La suggestibilité*. Paris: Schleicher frères. (Cit. on p. 74).

Binet, A. & Simon, T. (1916). *The development of intelligence in children* (E. Kite, Trans.). Baltimore: Williams & Wilkins company. (Cit. on p. 75).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley. (Cit. on pp. 41, 59, 92, 98).

Bollen, K. (1989). *Structural equations with latent variables.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley. (Cit. on pp. 96, 97).

Boring, E. G. (1920). The logic of the normal law of error in mental measurement. *The American journal of psychology*, *31*(1), 1–33. (Cit. on pp. 2, 60).

Boring, E. G. (1921). The Stimulus-Error. *The American Journal of Psychology*, *32*(4), 449–471. (Cit. on pp. 4, 22).

Boring, E. G. (1961). The Beginning and Growth Of Measurement in Psychology. In H. Woolf (Ed.), *Quantification: a history of the meaning measurement in the natural and social sciences* (pp. 108–127). Indianapolis: Bobbs-Merrill. (Cit. on p. 18).

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge; New York: Cambridge University Press. (Cit. on pp. 2, 10, 84, 111, 112, 123).

Borsboom, D. & Mellenbergh, G. J. (2004, January). Why psychometrics is not pathological. *Theory & Psychology*, *14*(1), 105–120. (Cit. on pp. 111, 113).

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. (Cit. on p. 84).

Boumans, M. (2001). Measure for measure: How economists model the world into numbers. *Social Research*, *68*(2), 427–453. (Cit. on p. 61).

Boumans, M. (2007). *Measurement in Economics: A Handbook*. London: Academic. (Cit. on p. 18).

Bourque, M. L. (2009). *A History of* NAEP *Achievement Levels: Issues, Implementation, and Impact 1989-2009*. Washington, DC: National Assessment Governing Board. (ED509389). (Cit. on pp. 6, 134).

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley. (Cit. on p. 64).

Brandom, R. (2003). *Articulating reasons: an introduction to inferentialism*. Cambridge: Harvard University Press. (Cit. on p. 14).

Brennan, R. L. (Ed.). (2006). *Educational measurement*. Westport, CT: Praeger Publishers. (Cit. on p. 18).

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillan Co. (Cit. on pp. 3, 19, 23–25).

Bureau International des Poids et Mesures. (2014). What is Metrology? Retrieved October 12, 2014, from http://www.bipm.org/en/worldwide-metrology/. (Cit. on p. 18)

Campbell, N. R. (1920). *Physics: the elements*. Cambridge: University Press. (Cit. on pp. 10, 26, 27, 31, 55, 67, 68, 78, 81).

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longmans, Green and Co. (Cit. on pp. 27, 110).

Campbell, N. R. (1940). Notes on Physical Measurement. *The Advancement of Science*, *1*(2), 340–342. (Cit. on p. 27).

Chang, H. (1995). Circularity and reliability in measurement. *Perspectives on Science*, *3*, 153–172. (Cit. on p. 60).

Chang, H. (2009). Operationalism. Retrieved October 12, 2014, from http://plato.stanford.edu/entries/operationalism/. (Cit. on p. 24)

Churchman, C. W. (1959). Why measure. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: definitions and theories*. London: John Wiley & Sons. (Cit. on p. 69).

Cizek, G. J. (2001). *Setting performance standards : concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates. (Cit. on pp. 133, 138).

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*(3), 186–190. (Cit. on pp. 2, 23, 33, 35–37, 115).

Cliff, N. & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates. (Cit. on p. 59).

Clogg, C. C. (1979). Some latent structure models for the analysis of likert-type data. *Social Science Research*, *8*(4), 287–301. (Cit. on p. 97).

Clogg, C. C. (1988). Latent class models for measuring. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 173–205). New York: Plenum Press. (Cit. on pp. 95, 99, 100).

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum Press. (Cit. on pp. 95, 99).

Clogg, C. C. & Goodman, L. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*(388), 762–771. (Cit. on p. 97).

Crocker, L. M. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston. (Cit. on p. 38).

Croon, M. (1990). Latent Class Analysis with Ordered Latent Classes. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 171–192. (Cit. on pp. 2, 6, 41, 103, 105, 133, 140, 141).

Croon, M. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 315–331. (Cit. on pp. 41, 105, 141).

Croon, M. (2002). Ordering the classes. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 137–162). New York: Cambridge University Press. (Cit. on pp. 41, 101, 105, 107, 133, 140, 141).

Dawes, R. & Smith, T. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology Vol. I* (pp. 509–566). Hillsdale, NJ: Lawrence Erlbaum. (Cit. on p. 35).

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press. (Cit. on pp. 38, 39).

De Boeck, P. & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer. (Cit. on pp. 2, 6, 95, 133).

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological review*, *112*(1), 129–158. (Cit. on p. 98).

De Leeuw, J. & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational and Behavioral Statistics*, *11*(3), 183–196. (Cit. on p. 99).

Dewey, J. (1960). *The quest for certainty: a study of the relation of knowledge and action*. New York: Putnam. (Original work published 1929). (Cit. on pp. 11, 15)

Dewey, J. (2004). *Reconstruction in philosophy*. Mineola, N.Y.: Dover Publications. (Original work published 1920). (Cit. on pp. 11, 16)

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science*, *1*(1), 5–26. (Cit. on pp. 10, 24, 53, 78, 79, 81, 82).

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19. (Cit. on p. 37).

Duncan, O. D. (1984). *Notes on social measurement: historical and critical*. New York: Russell Sage Foundation. (Cit. on pp. 1, 18, 32, 33, 44, 69, 77, 82).

Elmes, D. G., Kantowitz, B. H., & Roediger, H. L. (2012). *Research methods in psychology*. Australia: Wadsworth Cengage Learning. (Cit. on p. 33).

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175–186. (Cit. on p. 96).

Embretson, S. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum. (Cit. on p. 113).

Estes, W. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, *12*(3), 263–282. (Cit. on p. 36).

Euclid. (1956). *The thirteen books of Euclid's Elements* (T. L. Heath, Trans.). New York: Dover Publications. (Cit. on p. 20).

Everitt, B. (1984). *An introduction to latent variable models*. London: Chapman and Hall London. (Cit. on p. 94).

Falmagne, J.-C. (1976). Random conjoint measurement and loudness summation. *Psychological Review*, *83*(1), 65–79. (Cit. on p. 37).

Falmagne, J.-C. (1980). A Probabilistic Theory of Extensive Measurement. *Philosophy of Science*, *47*(2), 277–296. (Cit. on p. 37).

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*(1), 39–48. (Cit. on p. 43).

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359–374. (Cit. on pp. 2, 6, 133, 144).

Follmann, D. A. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika*, *53*(4), 553–562. (Cit. on p. 99).

Formann, A. (1982). Linear logistic latent class analysis. *Biometrical Journal*, *24*(2), 171–190. (Cit. on p. 103).

Formann, A. (1985). Constrained latent class models: Theory and applications. *British journal of mathematical & statistical psychology*, *38*(1), 87–111. (Cit. on p. 103).

Formann, A. (1989). Constrained latent class models: Some further applications. *British journal of mathematical & statistical psychology*, *42*(1), 37–54. (Cit. on p. 103).

Formann, A. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, *87*(418), 476–486. (Cit. on p. 103).

Formann, A. (1995). Linear logistic latent class analysis and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: foundations, recent developments, and applications* (pp. 239–256). New York: Springer-Verlag. (Cit. on pp. 41, 99, 114, 144).

Formann, A. & Kohlmann, T. (1998). Structural latent class models. *Sociological Methods and Research*, *26*(4), 530–565. (Cit. on pp. 99, 114).

Formann, A. & Thomas, K. (2002). Three-Parameter Linear Logistic Latent Class Analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 183–210). New York: Cambridge University Press. (Cit. on p. 139).

Franklin, W. (1903). Popular Science. *Science*, *17*(418), 8–15. (Cit. on p. 55).

Freund, R. (2014). *Scales, quantities, and locations*. Unpublished manuscript. (Cit. on p. 47).

Glaser, R. (1963). Instructional technology and the measurement of learing outcomes: Some questions. *American Psychologist*, *18*(8), 519–521. (Cit. on p. 135).

Glass, G. V. (1978). Standards and Criteria. *Journal of Educational Measurement*, *15*(4), 237–261. (Cit. on p. 138).

Goodman, L. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – A Modified latent structure approach. *American Journal of Sociology*, *79*(5), 1179–1259. (Cit. on p. 97).

Goodman, N. (1975). Words, works, worlds. *Erkenntnis*, 57–73. (Cit. on p. 64).

Goodman, R. B. (1998). Wittgenstein and Pragmatism. *Parallax*, *4*(4), 91–105. (Cit. on p. 13).

Gorsuch, R. L. (2003). Factor Analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology – Volume 2 – Research Methods in Psychology* (pp. 143–164). Hoboken, New Jersey: John Wiley & Sons, Inc. (Cit. on p. 97).

Grace, R. C. (2001). On the Failure of Operationism. *Theory & Psychology*, *11*(1), 5–33. (Cit. on p. 25).

Green, C. D. (1992). Of Immortal Mythological Beasts Operationism in Psychology. *Theory & Psychology*, *2*(3), 291–320. (Cit. on pp. 25, 32).

Green, C. D. (2001). Operationism again: What Did Bridgman Say? What Did Bridgman Need? *Theory & Psychology*, *11*(1), 45–51. (Cit. on p. 32).

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill. (Cit. on p. 38).

Guttman, L. (1950). The basis for scalogram analysis. In S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld, S. Star, & J. Clausen (Eds.), *Measurement and prediction* (Vol. 4, pp. 60–90). Princeton, N.J.: Princeton University Press. (Cit. on p. 98).

Haack, R. (1982). Wittgenstein's Pragmatism. *American Philosophical Quarterly*, *19*(2), 163–171. (Cit. on p. 13).

Haack, S. (1998). *Manifesto of a passionate moderate: unfashionable essays*. Chicago: University of Chicago Press. (Cit. on p. 14).

Haberman, S. J. (1979). *Analysis of qualitative data, vol. 2: New developments*. New York: Academic Press. (Cit. on p. 100).

Hacking, I. (2007). On not being a pragmatist: Eight reasons and a cause. In *New pragmatists* (pp. 32–49). New York: Oxford University Press. (Cit. on p. 67).

Hagenaars, J. A. & McCutcheon, A. L. (2002). *Applied latent class analysis*. New York: Cambridge University Press. (Cit. on pp. 41, 59, 139).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif.: Sage Publications. (Cit. on p. 140).

Hand, D. J. (2004). *Measurement theory and practice*. London: Arnold. (Cit. on pp. 10, 31, 73).

Heinen, T. (1993). *Discrete latent variable models*. Tilburg, The Netherlands: Tilburg University Press. (Cit. on pp. 94, 101, 102).

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage. (Cit. on pp. 39, 98, 100, 141).

Hensley, J. M. (2009). Who's Calling Wittgenstein a Pragmatist? *European Journal Of Pragmatism And American Philosophy*, *4*(2), 27–35. (Cit. on p. 13).

Hojtink, H. & Molenaar, I. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*(2), 171–189. (Cit. on pp. 103, 107, 140).

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. In *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig: Mathematische-Physische Classe* [Reports of the Proceedings of the Royal Saxon Society of Sciences in Leipzig: Mathematical-Physical Division] (Vol. 53, pp. 3–64). (Cit. on pp. 21, 47, 138).

Hornstein, G. A. (1988). Quantifying psychological phenomena: Debates, dilemmas, and implications. In J. G. Morawski (Ed.), *The rise of experimentation in american psychology* (pp. 1–34). New Haven: Yale University Press. (Cit. on pp. 23, 44).

Huntington, E. V. (1902). A complete set of postulates for the theory of absolute continuous magnitude. *Transactions of the American Mathematical Society*, *3*(2), 264–279. (Cit. on pp. 21, 138).

Hutchins, E. (2000). *Cognition in the wild*. Cambridge, Massachusetts: The MIT Press. (Cit. on p. 15).

International Measurement Confederation. (2014). Table of Technical Committees. Retrieved October 12, 2014, from http://www.imeko.org. (Cit. on p. 18)

James, W. (1995). *Pragmatism*. New York: Dover Publications. (Original work published 1907). (Cit. on pp. 1, 2, 11, 12, 14, 15, 55)

Jensen, M. P. & Karoly, P. (1992). Self-report scales and procedures for assessing pain in adults. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (pp. 19–41). New York: Guilford Press. (Cit. on p. 138).

Johnson, H. M. (1936). Pseudo-mathematics in the mental and social sciences. *The American journal of psychology*, *48*(2), 342–351. (Cit. on pp. 2, 4, 9, 23, 45).

Joint Committee for Guides in Metrology. (2008). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. Paris, France: Bureau International des Poids et Mesures. (Cit. on p. 41).

Joint Committee for Guides in Metrology. (2012). *The international vocabulary of metrology – basic and general concepts and associated terms (*VIM*)* (3rd ed.). Paris, France: Bureau International des Poids et Mesures. (Cit. on pp. 41, 46, 60, 80–83).

Jöreskog, K. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. (Cit. on p. 97).

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, *56*(2), 255–278. (Cit. on p. 96).

Kaarls, R. (2007). *Evolving Needs for Metrology in Trade, Industry and Society and the Role of the* BIPM. Paris: Bureau International des Poids et Mesures. (Cit. on p. 41).

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *49*(2), 223–245. (Cit. on p. 100).

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. Psychology Press. (Cit. on p. 113).

Knapp, T. R. & Mueller, R. O. (2010). Reliability and validity of instruments. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 337–341). New York, NY: Routledge. (Cit. on p. 8).

Koch, S. (1992). Psychology's Bridgman vs Bridgman's Bridgman: An Essay in Reconstruction. *Theory and Psychology*, *2*(3), 261–290. (Cit. on p. 25).

Krantz, D. H. (1967). Extensive Measurement in Semiorders. *Philosophy of Science*, *34*(4), 348–362. (Cit. on p. 37).

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (2007). *Foundations of measurement. Volume I: Additive and polynomial representations*. Mineola, NY: Dover Publications Inc. (Original work published 1971). (Cit. on pp. 3, 10, 19, 26, 33, 45, 83, 110)

Kuhn, T. S. (1961). The function of measurement in modern physical science. *ISIS*, *52*(2), 161–193. (Cit. on p. 60).

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89–109. (Cit. on p. 113).

Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh : the embodied mind and its challenge to Western thought*. New York: Basic Books. (Cit. on p. 66).

Langeheine, R. & Rost, J. (1988). *Latent trait and latent class models*. New York: Plenum Press. (Cit. on pp. 39, 98, 100).

Laurence, S. & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: core readings* (pp. 3–81). Cambridge, Mass: MIT Press. (Cit. on pp. 16, 48, 49).

Lawley, D. & Maxwell, A. (1971). *Factor analysis as a statistical method*. New York: American Elsevier Pub. Co. (Cit. on p. 97).

Lazarsfeld, P. F. (1961). Notes on the History Of Quantification in Sociology – Trends, Sources And Problems. In H. Woolf (Ed.), *Quantification: a history of the meaning measurement in the natural and social sciences* (pp. 147–203). Indianapolis: Bobbs-Merrill. (Cit. on p. 18).

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin Company. (Cit. on pp. 3, 19, 39, 41, 59, 94, 95, 98, 139).

Lehrer, R., Kim, M., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In J. Confrey & A. Maloney (Eds.), *Learning over time: learning trajectories in mathematics education* (pp. 31–60). Charlotte, NC: Information Age Publishers. (Cit. on p. 7).

Lehrer, R. & Wilson, M. (2011, April). *Developing assessments of data modeling: Construct maps as boundary objects*. Paper presented at the annual meeting of the annual meeting of the American Educational Research Association, New Orleans, LA. (Cit. on p. 135).

Ligtvoet, R. (2010). *Essays on invariant item ordering*. (Unpublished doctoral dissertation). Tilburg University, The Netherlands. (Cit. on p. 126).

Lindquist, E. F. & Thorndike, R. L. (Eds.). (1951). *Educational measurement*. Washington: American Council on Education. (Cit. on p. 18).

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96–107. (Cit. on pp. 99, 114, 139).

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, *61*(4), i–49. (Cit. on p. 107).

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin*, *45*(6), 507–529. (Cit. on p. 107).

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Pub. Co. (Cit. on pp. 3, 19, 38, 73, 98).

Lubke, G. & Muthen, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. (Cit. on pp. 41, 139).

Luce, R. D. (1956). Semiorders and a Theory of Utility Discrimination. *Econometrica*, *24*(2), 178–191. (Cit. on p. 37).

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (2007). *Foundations of measurement. Volume III*. Mineola, NY: Dover Publications Inc. (Original work published 1990). (Cit. on p. 33)

Luce, R. D. & Suppes, P. (2002). Representational measurement theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology* (Vol. 4, pp. 1–41). New York: John Wiley and Sons. (Cit. on p. 78).

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27. (Cit. on pp. 33, 35, 110).

Margolis, E. & Laurence, S. (2014). Concepts. Retrieved October 12, 2014, from http://plato.stanford.edu/archives/spr2014/entries/concepts/. (Cit. on pp. 15, 16)

Margolis, J. (2009). A Philosophical Bestiary. *European Journal of Pragmatism & American Philosophy*, *4*(2), 128–145. (Cit. on p. 13).

Mari, L. (2000). Beyond the representational viewpoint: a new formalization of measurement. *Measurement*, *27*(2), 71–84. (Cit. on p. 10).

Mari, L. (2003). Epistemology of measurement. *Measurement*, *34*(1), 17–30. (Cit. on pp. 10, 61).

Mari, L. (2013). A quest for the definition of measurement. *Measurement*, *46*(8), 2889–2895. (Cit. on pp. 43, 47).

Mari, L. (2014). Evolution of 30 years of the International Vocabulary of Metrology (VIM). *Metrologia*, *52*(1), R1–R10. (Cit. on p. 42).

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. (Cit. on pp. 96, 132).

Markus, K. A. & Borsboom, D. (2013). *Frontiers of test validity theory : measurement, causation and meaning*. New York: Routledge. (Cit. on p. 24).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. (Cit. on p. 99).

Masters, G. N. (1985). A comparison of latent trait and latent class analyses of Likert-type data. *Psychometrika*, *50*(1), 69–82. (Cit. on pp. 95, 98, 99).

Maul, A. (2014). Justification Is Not Truth, and Testing Is Not Measurement: Understanding the Purpose and Limitations of the Standards. *Educational Measurement: Issues and Practice*, *33*(4), 39–41. (Cit. on p. 80).

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, California: Sage Publications, Inc. (Cit. on pp. 96, 101).

McDonald, R. P. (1999). *Test theory a unified treatment*. Mahwah, N.J.: Lawrence Erlbaum Associates. (Cit. on p. 38).

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*(12), 1469–1481. (Cit. on p. 49).

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283–298. (Cit. on pp. 106, 111, 129).

Menand, L. (2001). *The Metaphysical Club*. New York: Farrar, Straus, and Giroux. (Cit. on p. 17).

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–104). New York, NY: Macmillan Publishing Company. (Cit. on pp. 38, 79).

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc. (Cit. on pp. 20, 31, 33, 78, 84, 108–110, 138).

Michell, J. (1993). The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell. *Studies in History and Philosophy of Science Part A*, *24*(2), 185–206. (Cit. on pp. 26, 28).

Michell, J. (1997a). Bertrand Russell's 1897 critique of the traditional theory of measurement. *Synthese*, *110*(2), 257–276. (Cit. on p. 26).

Michell, J. (1997b). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. (Cit. on pp. 18, 20–22, 45, 55, 81).

Michell, J. (1999). *Measurement in psychology: critical history of a methodological concept*. New York: Cambridge University Press. (Cit. on pp. 1, 4, 10, 20–23, 26, 28, 33, 44, 45, 84, 86, 108, 109).

Michell, J. (2003). Pragmatism, positivism and the quantitative imperative. *Theory & Psychology*, *13*(1), 45–52. (Cit. on p. 71).

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, *38*(4), 285–294. (Cit. on pp. 3, 10, 19–22, 45, 84, 86, 109).

Michell, J. (2008a). Conjoint Measurement and the Rasch Paradox. *Theory & Psychology*, *18*(1), 119–124. (Cit. on pp. 108, 113).

Michell, J. (2008b). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1), 7–24. (Cit. on pp. 2, 4, 108, 109, 115, 131).

Michell, J. (2008c). The Measure of Psychometrics — Review: Denny Borsboom, Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. *Theory & Psychology*, *18*(1), 135–137. (Cit. on pp. 68, 78).

Michell, J. (2011). Qualitative research meets the ghost of Pythagoras. *Theory & Psychology*, *21*(2), 241–259. (Cit. on pp. 20, 22, 65).

Michell, J. (2012a). Alfred Binet and the concept of heterogeneous orders. *Frontiers in psychology*, *3*(261), 1–8. (Cit. on pp. 21, 46, 74, 75).

Michell, J. (2012b). "The constantly recurring argument": Inferring quantity from order. *Theory and Psychology*, *22*(3), 255–271. (Cit. on p. 45).

Michell, J. & Ernst, C. (1996). The axioms of quantity and the theory of measurement: Translated from Part I of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of mathematical psychology*, *40*(3), 235–252. (Cit. on pp. 21, 47).

Michell, J. & Ernst, C. (1997). The axioms of quantity and the theory of measurement: Translated from Part II of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of mathematical psychology*, *41*(4), 345–356. (Cit. on pp. 21, 47).

Minsky, M. (1968). Matter, Mind and Models. In M. Minsky (Ed.), *Semantic information processing* (pp. 425–432). Cambridge, Mass: MIT Press. (Cit. on p. 64).

Misak, C. J. (2013). *The American pragmatists*. Oxford: Oxford University Press. (Cit. on p. 14).

Mislevy, R. J. (1996). Test Theory Reconceived. *Journal of Educational Measurement*, *33*(4), 379–416. (Cit. on pp. 6, 135).

Mislevy, R. J. & Haertel, G. (2006). Implications of Evidence Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. (Cit. on pp. 79, 124).

Mislevy, R. J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215. (Cit. on pp. 2, 6, 148).

Mohs, F. (1825). *Treatise on Mineralogy, or the Natural History of the Mineral Kingdom* (W. Haidinger, Trans.). Edinburgh: Constable and Co. (Cit. on p. 138).

Mokken, R. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague: Mouton. (Cit. on pp. 98, 103).

Mokken, R. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*(4), 417–430. (Cit. on p. 103).

Moustaki, I. & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411. (Cit. on p. 100).

Muthen, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117. (Cit. on p. 100).

Muthen, B. O. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, *31*(6), 1050–1066. (Cit. on pp. 41, 139).

Nagel, E. (1931). Measurement. *Erkenntnis*, *2*(1), 313–333. (Cit. on p. 26).

Narens, L. (2013). *Introduction to the theories of measurement and meaningfulness and the use of symmetry in science*. Hove: Psychology Press. (Cit. on pp. 26, 33).

Narens, L. & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166–180. (Cit. on pp. 33–35, 78, 82).

Narens, L. & Luce, R. D. (1993). Further comments on the "nonrevolution" arising from axiomatic measurement theory. *Psychological Science*, *4*(2), 127–130. (Cit. on pp. 35, 37).

National Center for Education Statistics. (2009). *Highlights from* TIMSS *2007 mathematics and science achievement of* U.S. *fourth- and eighth-grade students in an international context*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education. (ED503625). (Cit. on pp. 6, 134).

National Research Council. (2000). *How people learn : brain, mind, experience, and school*. Washington, D.C.: National Academy Press. (Cit. on p. 135).

National Research Council. (2001). *Knowing what students know : the science and design of educational assessment*. Washington, DC: National Academy Press. (Cit. on pp. 6, 135).

National Research Council. (2006). *Systems for state science assessment*. Washington, DC: National Academies Press. (Cit. on pp. 7, 135).

National Research Council. (2007). *Taking science to school : learning and teaching science in grades K-8*. Washington, D.C.: National Academies Press. (Cit. on p. 135).

Nature Publishing Group. (1939). Quantitative Estimates of Sensory Events. *Nature*, *144*(3658), 973–973. (Cit. on p. 29).

Nestor, P. G. & Schutt, R. K. (2014). *Research methods in psychology: Investigating human behavior*. Thousand Oaks, CA: Sage Publications. (Cit. on p. 33).

Neurath, O. (1973). Anti-Spengler. In M. Neurath & R. S. Cohen (Eds.), *Empiricism and sociology* (pp. 158–213). Dordrecht, Holland: D. Reidel Publishing Company. (Cit. on p. 2).

Núñez, R. E. & Freeman, W. J. (Eds.). (1999). *Reclaiming cognition: the primacy of action, intention, and emotion*. Thorverton, UK: Imprint Academic. (Cit. on p. 15).

OECD. (2007). PISA *2006 : science competencies for tomorrow's world*. Paris: OECD. (Cit. on pp. 6, 134).

OECD. (2014). PISA *2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. París: OECD. (Cit. on p. 134).

Peirce, C. S. (1868). Some Consequences Of Four Incapacities. *The Journal of Speculative Philosophy*, *2*(3), 140–157. (Cit. on pp. 13, 14).

Peirce, C. S. (2014). How to make our ideas clear. In J. Buchler (Ed.), *Philosophical writings of Peirce* (pp. 23–41). New York: Dover Publications. (Original work published 1878). (Cit. on pp. 11, 12, 108)

Perie, M. (2008). A Guide to Understanding and Developing Performance-Level Descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29. (Cit. on pp. 6, 134).

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255. (Cit. on pp. 37, 111, 113).

Pfanzagl, J. (1968). *Theory of measurement*. New York: Wiley. (Cit. on pp. 3, 19, 26, 33, 45).

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press. (Cit. on pp. 9, 43).

Psychometric Society. (2014). Psychometrics & the Psychometric Society. Retrieved October 14, 2014, from https://www.psychometricsociety.org/content/psychometrics-psychometric-society. (Cit. on p. 18)

Putnam, H. (1995). *Pragmatism: an open question*. Cambridge, Mass., USA: Blackwell Publishers Inc. (Cit. on p. 13).

Quine, W. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*(1), 20–43. (Cit. on p. 124).

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. (Cit. on p. 153).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Probability theory*, 321. (Cit. on pp. 99, 112, 113).

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. In *European meeting on statistics, econometrics and management science, amsterdam* (Vol. 2, 7). (Cit. on pp. 99, 112).

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Original work published 1960). (Cit. on pp. 3, 19, 41, 59, 79, 92, 98, 99, 111, 113, 144)

Reckase, M. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36. (Cit. on p. 96).

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer Verlag. (Cit. on p. 96).

Rescher, N. (2007). Pragmatism. In C. V. Boundas (Ed.), *Columbia companion to twentieth-century philosophies* (pp. 128–142). New York: Columbia University Press. (Cit. on pp. 12, 14).

Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton: Princeton University Press. (Cit. on p. 13).

Rorty, R. (1982). *Consequences of pragmatism: essays, 1972-1980*. Minneapolis: University of Minnesota Press. (Cit. on pp. 11, 14).

Rorty, R. (1999). *Philosophy and social hope*. London, UK: Penguin. (Cit. on pp. 1, 2, 11, 14, 15, 63, 66).

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 24–78). Hilldate, NJ: Lawrence Erlbaum. (Cit. on pp. 15, 48).

Rosch, E. (1999). Reclaiming concepts. *Journal of consciousness studies*, *6*(11-12), 61–77. (Cit. on pp. 15, 16, 48–50).

Rosch, E. & Lloyd, B. (1978). *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates. (Cit. on p. 15).

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. (Cit. on pp. 16, 49).

Rost, J. (1988). Test theory with qualitative and quantitative latent variables. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 147–171). New York: Plenum Press. (Cit. on pp. 98, 126).

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. (Cit. on pp. 96, 126).

Rost, J. & Langeheine, R. (1997). *Applications of latent trait and latent class models in the social sciences*. New York: Waxmann. (Cit. on p. 98).

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement : theory, methods, and applications*. New York: Guilford Press. (Cit. on pp. 96, 132, 140).

Russell, B. (1897). On the Relations of Number and Quantity. *Mind*, *6*(23), 326–341. (Cit. on p. 26).

Russell, B. (1903). *The principles of mathematics, vol. 1*. Cambridge: University Press. (Cit. on p. 26).

Savage, C. W. (1970). *The measurement of sensation; a critique of perceptual psychophysics*. Berkeley: University of California Press. (Cit. on pp. 10, 44).

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models (ASISOP). *Psychometrika*, *64*(3), 295–316. (Cit. on pp. 111, 113).

Schönemann, P. H. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. *Trends and perspectives in empirical social research*, 149–160. (Cit. on pp. 2, 4, 34, 36, 109, 115, 131).

Schwager, K. W. (1991). The representational theory of measurement: An assessment. *Psychological Bulletin*, *110*(3), 618–626. (Cit. on p. 36).

Scott, D. & Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logic*, *23*(2), 113–128. (Cit. on pp. 33, 34).

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science, vol. i* (pp. 253–329). Minneapolis, MN: University of Minnesota Press. (Cit. on pp. 13, 14).

Shanon, B. (1993). *The representational and the presentational: an essay on cognition and the study of mind*. London: Harvester Wheatsheaf. (Cit. on p. 15).

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, *42*(4), 509–524. (Cit. on p. 76).

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*(1), 3–31. (Cit. on p. 103).

Sijtsma, K. & Molenaar, I. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, California: Sage Publications, Inc. (Cit. on pp. 104, 107).

Simpson, G. (1906). *The Beaufort Scale of Wind-Force* (tech. rep. No. 180). The Meteorological Office. London, UK. (Cit. on p. 138).

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC. (Cit. on pp. 38, 39, 61, 95, 98, 100).

Smith, C., Wiser, M., Anderson, C., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research & Perspectives*, *14*(1-2), 1–98. (Cit. on pp. 7, 135).

Spearman, C. (1904a). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. (Cit. on pp. 38, 96).

Spearman, C. (1904b). The proof and measurement of association between two things. *The American journal of psychology*, *15*(1), 72–101. (Cit. on p. 38).

Speitel, K. (1992). Measurement assurance. In G. Salvendy (Ed.), *Handbook of industrial engineering* (pp. 2235–2251). New York: Wiley. (Cit. on p. 10).

Spengler, J. J. (1961). On the Progress Of Quantification in Economics. In H. Woolf (Ed.), *Quantification: a history of the meaning measurement in the natural and social sciences* (pp. 128–146). Indianapolis: Bobbs-Merrill. (Cit. on p. 18).

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. (Cit. on pp. 10, 27, 29, 30, 34, 38, 45, 47, 56, 81, 82).

Stevens, S. S. (1958). Measurement and Man. *Science*, *127*(3295), 383–389. (Cit. on p. 31).

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617. (Cit. on pp. 96, 103).

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement*, *25*(3), 300–306. (Cit. on p. 103).

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*(4), 485–518. (Cit. on p. 96).

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (2007). *Foundations of measurement. Volume II*. Mineola, NY: Dover Publications Inc. (Original work published 1989). (Cit. on p. 33)

Suppes, P. & Zinnes, J. L. (1962). *Basic Measurement Theory* (tech. rep. No. 45). Institute For Mathematical Studies In The Social Sciences - Stanford University. Stanford, California. (Cit. on p. 83).

Suppes, P. & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, vol. 1* (pp. 3–76). New York: Wiley. (Cit. on pp. 33, 34).

Tal, E. (2013). Old and New Problems in Philosophy of Measurement. *Philosophy Compass*, *8*(12), 1159–1173. (Cit. on p. 60).

Tal, E. (2015). Measurement in Science. Retrieved July 28, 2015, from http://plato.stanford.edu/entries/measurement-science/. (Cit. on pp. 84, 85)

Tarpey, T. (2009). *All Models are Right… most are useless*. http://andrewgelman.com/wp-content/uploads/2012/03/tarpey.pdf. (Cit. on p. 64).

Tartaglia, J. (2007). *Routledge philosophy guidebook to Rorty and the mirror of nature*. London; New York: Routledge. (Cit. on pp. 17, 64, 69).

Teller, P. (2001). Twilight Of The Perfect Model Model. *Erkenntnis*, *55*(3), 393–415. (Cit. on p. 61).

Thurstone, L. L. (1947). *Multiple-factor analysis*. University of Chicago Press. (Cit. on p. 97).

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, New York: Wiley. (Cit. on p. 139).

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*(1), 23–30. (Cit. on p. 100).

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley. (Cit. on pp. 32, 38, 82).

Trendler, G. (2009). Measurement Theory, Psychology and the Revolution That Cannot Happen. *Theory and Psychology*, *19*(5), 579–599. (Cit. on pp. 2, 4, 9, 109, 131).

Trout, J. D. (1998). *Measuring the intentional world realism, naturalism, and quantitative methods in the behavioral sciences*. New York: Oxford University Press. (Cit. on p. 56).

Tucker, L. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, *11*(1), 1–13. (Cit. on p. 97).

Uebersax, J. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, *88*(422), 421–427. (Cit. on pp. 41, 98, 99, 139).

van der Linden, W. J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: theory into practice* (Vol. 2, pp. 3–24). Norwood, N.J.: Ablex Pub. Corp. (Cit. on pp. 39, 110).

van der Linden, W. J. (1995). A conceptual analysis of standard-setting in large-scale assessments. In *Proceedings of the joint conference on standard-setting for large-scale assessments* (Vol. 2, pp. 97–118). (Cit. on p. 134).

Van Fraassen, B. C. (2008). *Scientific representation: paradoxes of perspective*. New York: Oxford University Press. (Cit. on pp. 8, 60).

Van Onna, M. (2004). *Ordered latent class models in nonparametric item response theory*. (Unpublished doctoral dissertation). University of Groningen, The Netherlands. (Cit. on pp. 41, 96, 107, 140).

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: cognitive science and human experience*. Cambridge, Mass.: MIT Press. (Cit. on pp. 15, 66, 67).

Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72. (Cit. on p. 33).

Vermunt, J. K. & Magidson, J. (2005). *Latent* GOLD *4.0 user's guide*. Belmont, Massachusetts: Statistical Innovations Inc. (Cit. on pp. 153, 160).

Vermunt, J. K. & Magidson, J. (2008). *LG-syntax user's guide: Manual for Latent* GOLD *4.5 syntax module*. Belmont, Massachusetts: Statistical Innovations Inc. (Cit. on pp. 153, 160).

Weitzenhoffer, A. M. (1951). Mathematical structures and psychological measurements. *Psychometrika*, *16*(4), 387–406. (Cit. on p. 38).

Westbrook, R. (2008). The Pragmatist Family Romance. In C. Misak (Ed.), *The Oxford Handbook of American Philosophy* (pp. 185–196). Oxford: Oxford University Press. (Cit. on p. 11).

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics*, *13*(1), 1–14. (Cit. on p. 9).

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276–289. (Cit. on p. 96).

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah, N.J.: Lawrence Erlbaum Associates. (Cit. on pp. 6, 38, 79, 124, 135).

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730. (Cit. on pp. 7, 135).

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, *46*(9), 3766–3774. (Cit. on p. 61).

Wittgenstein, L. (1965). *Preliminary studies for the "Philosophical investigations", generally know as the Blue and Brown books*. New York: Harper and Row. (Cit. on p. 13).

Wittgenstein, L. (2003). *Philosophical investigations: the German text, with a Revised English translation* (G. E. M. Anscombe, Trans.). Malden, MA: Blackwell Pub. (Original work published 1953). (Cit. on pp. 12, 13, 49, 50)

Wolins, L. (1978). Interval Measurement: Physics, Psychophysics, and Metaphysics. *Educational and Psychological Measurement*, *38*(1), 1–9. (Cit. on p. 33).

Woolf, H. (Ed.). (1961). *Quantification: a history of the meaning measurement in the natural and social sciences*. Indianapolis: Bobbs-Merrill. (Cit. on p. 1).

Wright, B. D. (1997). A history of social science measurement. *Educational measurement: issues and practice*, *16*(4), 33–45. (Cit. on p. 79).

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press. (Cit. on p. 138).

Zieky, M. J. (1995). A historical perspective on setting standards. In *Proceedings of the joint conference on standard setting for large-scale assessments* (Vol. 2, pp. 1–37). Rockvile, MD.: Aspen Systems Corp. (Cit. on p. 133).